Technical Report Documentation Page

1. REPORT No.

2. GOVERNMENT ACCESSION No.

3. RECIPIENT'S CATALOG No.

4. TITLE AND SUBTITLE

Applications Of Statistics In Analyzing Aerometric Data For Transportation Systems

5. REPORT DATE

October 1974

6. PERFORMING ORGANIZATION

7. AUTHOR(S)

Gerald R. Bemis

8. PERFORMING ORGANIZATION REPORT No.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

State of California
Department of Transportation
Division of Highways
Transportation Laboratory

10. WORK UNIT No.

11. CONTRACT OR GRANT No.

13. TYPE OF REPORT & PERIOD COVERED

12. SPONSORING AGENCY NAME AND ADDRESS

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

The information in this manual consists of lecture notes for an Air Quality Training Course given to the Transportation Districts for the purpose of applying statistical analyses techniques to assess the impact of transportation systems on the environment.

16. ABSTRACT

A number of studies must be completed prior to the writing of an environmental impact statement for a highway project. One of these studies is concerned with the gathering of air quality field data, analysis of such data, and the writing of an air quality report.

The California Department of Transportation has embarked on a program of equipping and training district personnel to prepare air quality reports. This requires a two-week training course and the preparation of air quality manuals to be used as guides in the gathering of field data, analysis of results, and report writing.

17. KEYWORDS

18. No. OF PAGES:

377

19. DRI WEBSITE LINK

http://www.dot.ca.gov/hq/research/researchreports/1974-1975/74-38.pdf

20. FILE NAME

74-38.pdf

STATE OF CALIFORNIA
DEPARTMENT OF TRANSPORTATION
DIVISION OF HIGHWAYS
TRANSPORTATION LABORATORY

October 1974

Mr. R. J. Datel
State Highway Engineer

Dear Sir:

I have reviewed and now submit for your information this final research project report titled:

APPLICATIONS OF STATISTICS IN ANALYZING
AEROMETRIC DATA FOR TRANSPORTATION SYSTEMS

Study made by	Environmental Improvement Section
Under the Supervision of...........	Earl C. Shirley, P. E.
Principal Investigator.............	Andrew J. Ranzieri, P. E.
Report Prepared by.................	Gerald R. Bemis, P. E.
Assisted by........................	Paul D. Allen, P. E. Steve Morse, P. E. and James S. L. Fong

Very truly yours,

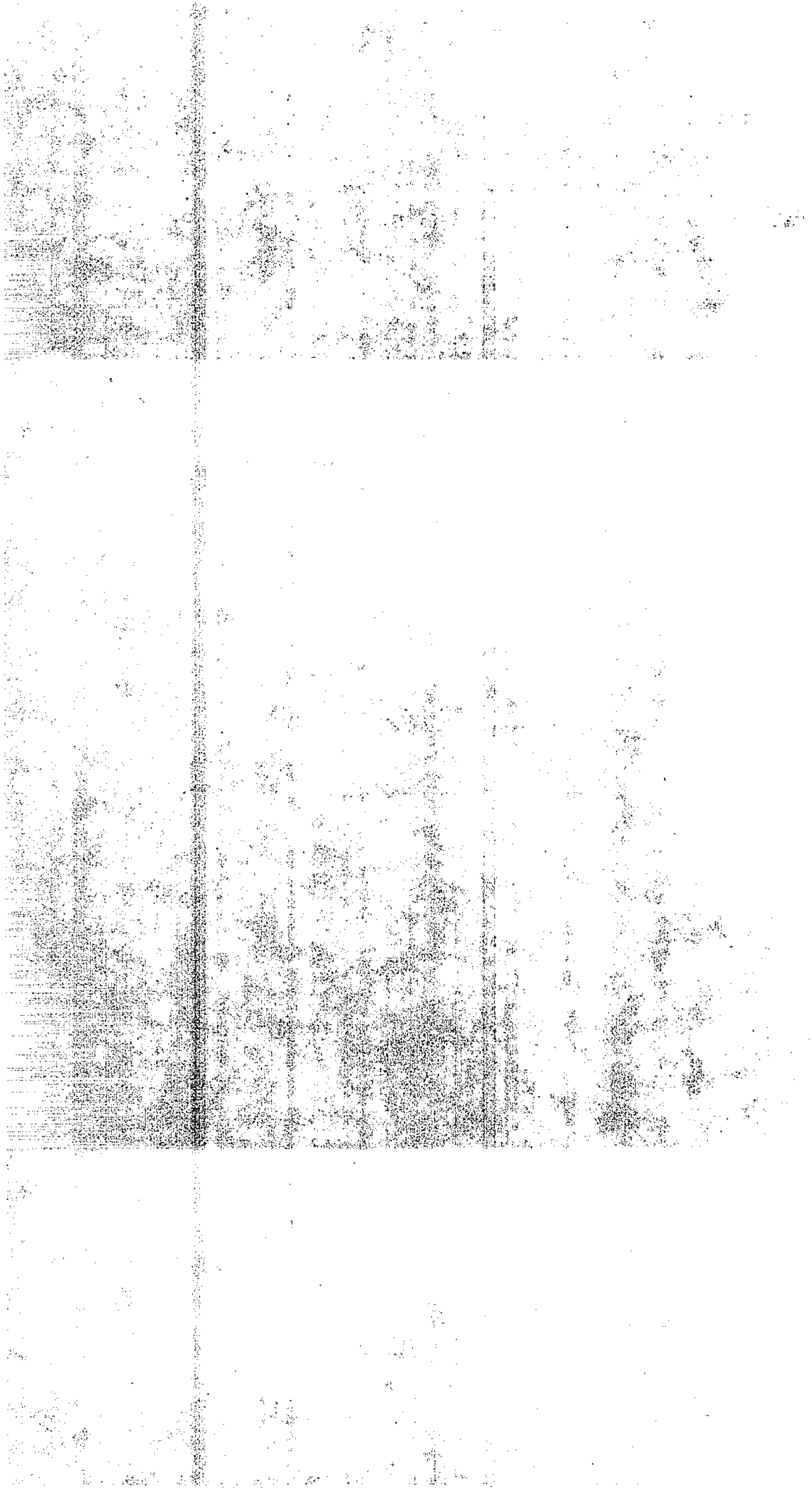
JOHN L. BEATON
Chief Engineer, Transportation Laboratory

Attachment

74-38

The contents of this report reflect the views of the Transportation Laboratory which is responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation.

The information in this manual consists of lecture notes for an Air Quality Training Course given to the Transportation Districts for the purpose of applying statistical analyses techniques to assess the impact of transportation systems on the environment.



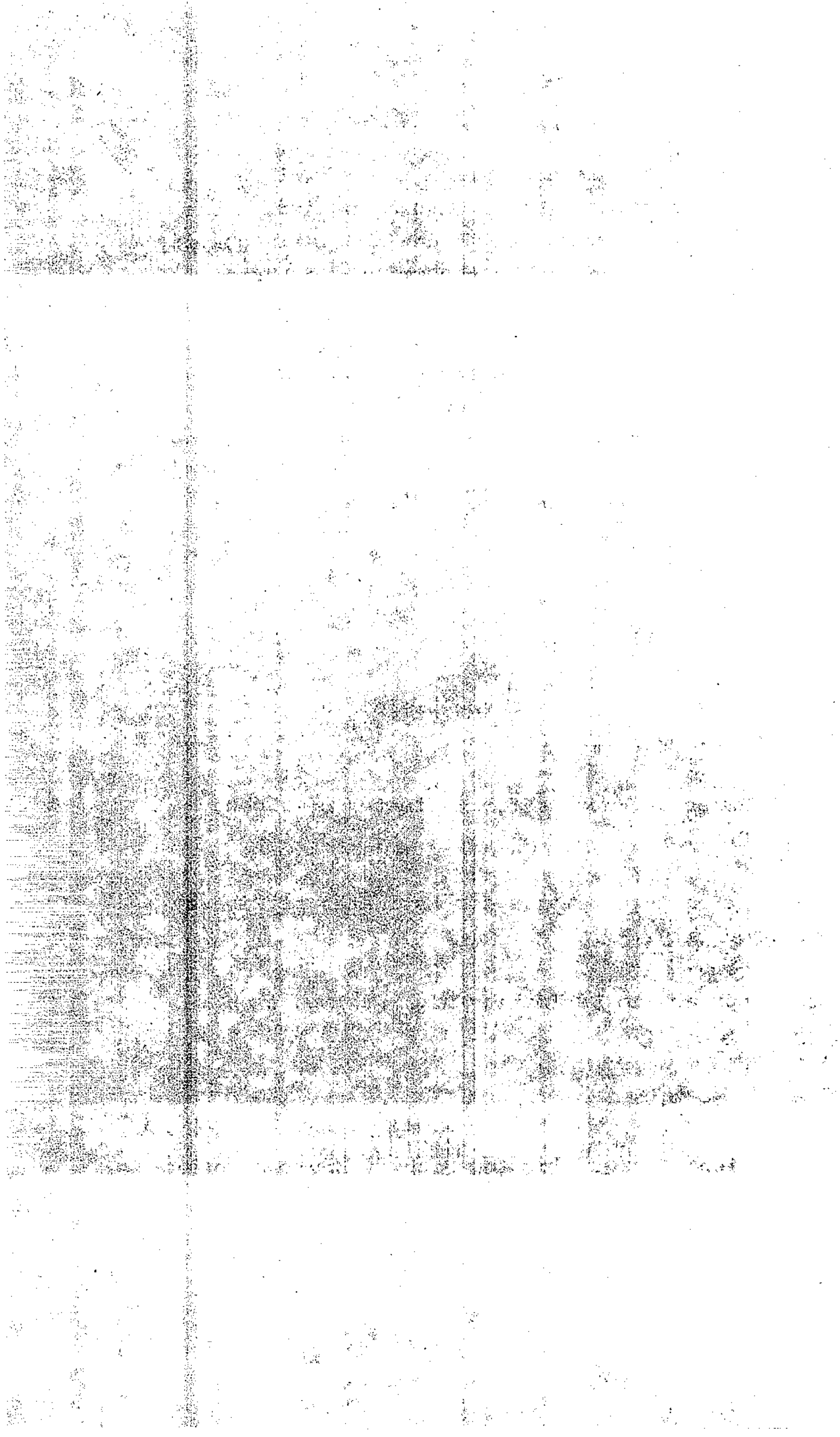
FOREWORD

A number of studies must be completed prior to the writing of an environmental impact statement for a highway project. One of these studies is concerned with the gathering of air quality field data, analysis of such data, and the writing of an air quality report.

The California Department of Transportation has embarked on a program of equipping and training district personnel to prepare air quality reports. This requires a two-week training course and the preparation of air quality manuals to be used as guides in the gathering of field data, analysis of results, and report writing.

This manual is the tenth in a series of air quality manuals, the titles of which are the following:

1. Meteorology and Its Influence on the Dispersion of Pollutants from Highway Line Sources.
2. Motor Vehicle Emission Factors for Estimates of Highway Impact on Air Quality.
3. Traffic Information Requirements for Estimates of Highway Impact on Air Quality.
4. Mathematical Approach to Estimating Highway Impact on Air Quality.
5. Appendix to Volume 4.
6. Analysis of Ambient Air Quality for Highway Environmental Projects.



7. A Method for Analyzing and Reporting Highway Impact on Air Quality.

8. Synthesis of Information on Highway Transportation and Air Quality.

9. Applications of Regression Analysis to Environmental Problems for Highway Projects.

10. Applications of Statistics in Analyzing Aerometric Data for Transportation Systems.

It is assumed that the reader of this manual is familiar with all of the above manuals and has a background in statistics.

At the end of this manual are listings of computer programs for all statistical tests discussed.

ACKNOWLEDGEMENTS

This manual has been authored by Andrew J. Ranzieri and Gerald R. Bemis under the supervision of Earl C. Shirley, Chief, Environmental Improvement Section.

Special acknowledgement is given to Paul Allen of the Transportation Laboratory for his work in developing the frequency distribution method used in Larsen's Model applications and the computer program which uses this method.

Also acknowledged is the assistance of James Fong who prepared the basic introduction, and Steve Morse who wrote part of the text for Larsen's Model. They are employed by the California Department of Transportation Environmental Planning Branch, District 04, San Francisco, California.

Further acknowledgement is given to Charles Frazier and Dil Onodera of the Transportation Laboratory in Sacramento for their assistance in developing the computer programs.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

EXAMPLES OF REGRESSION ANALYSES

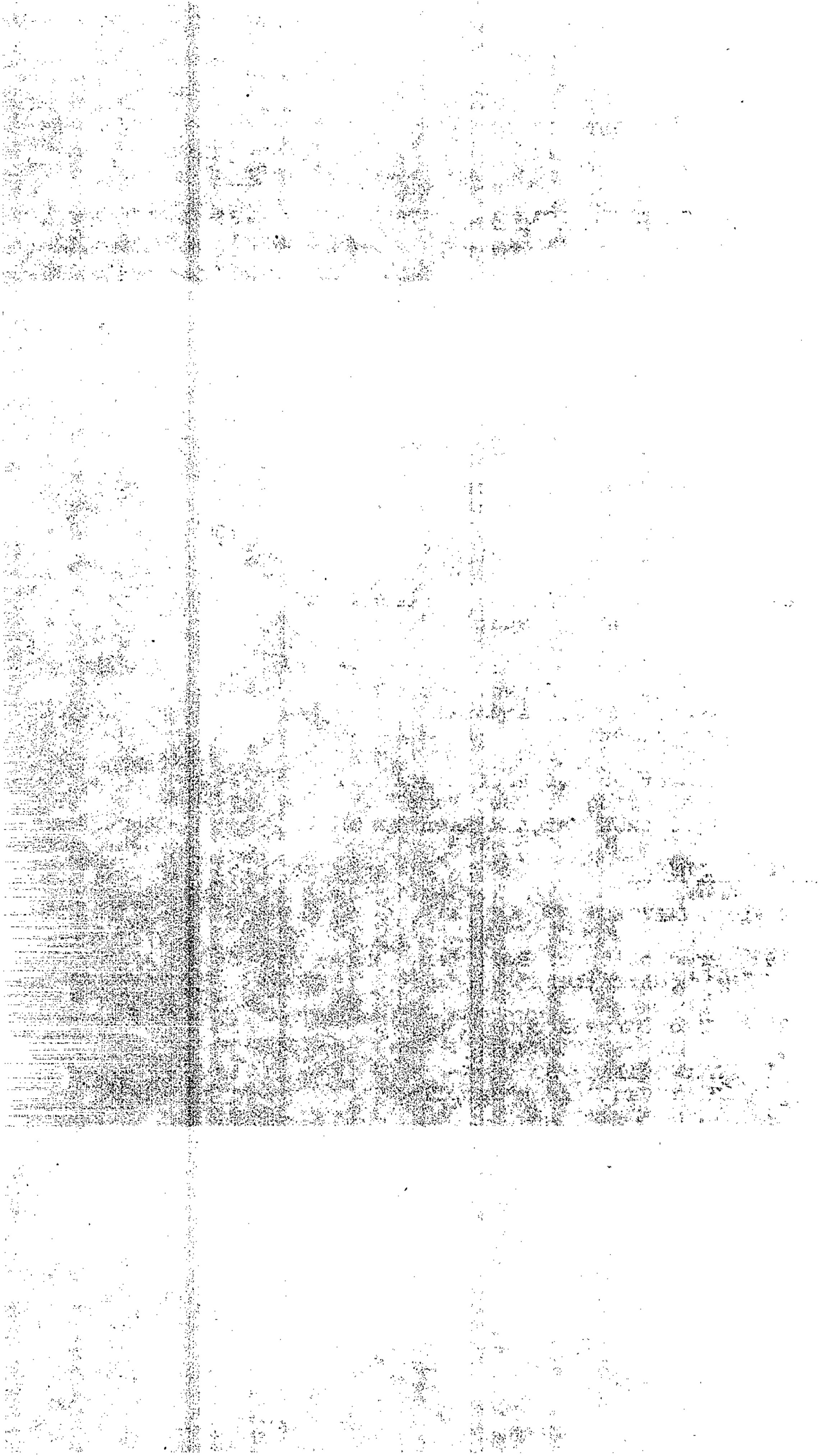


TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

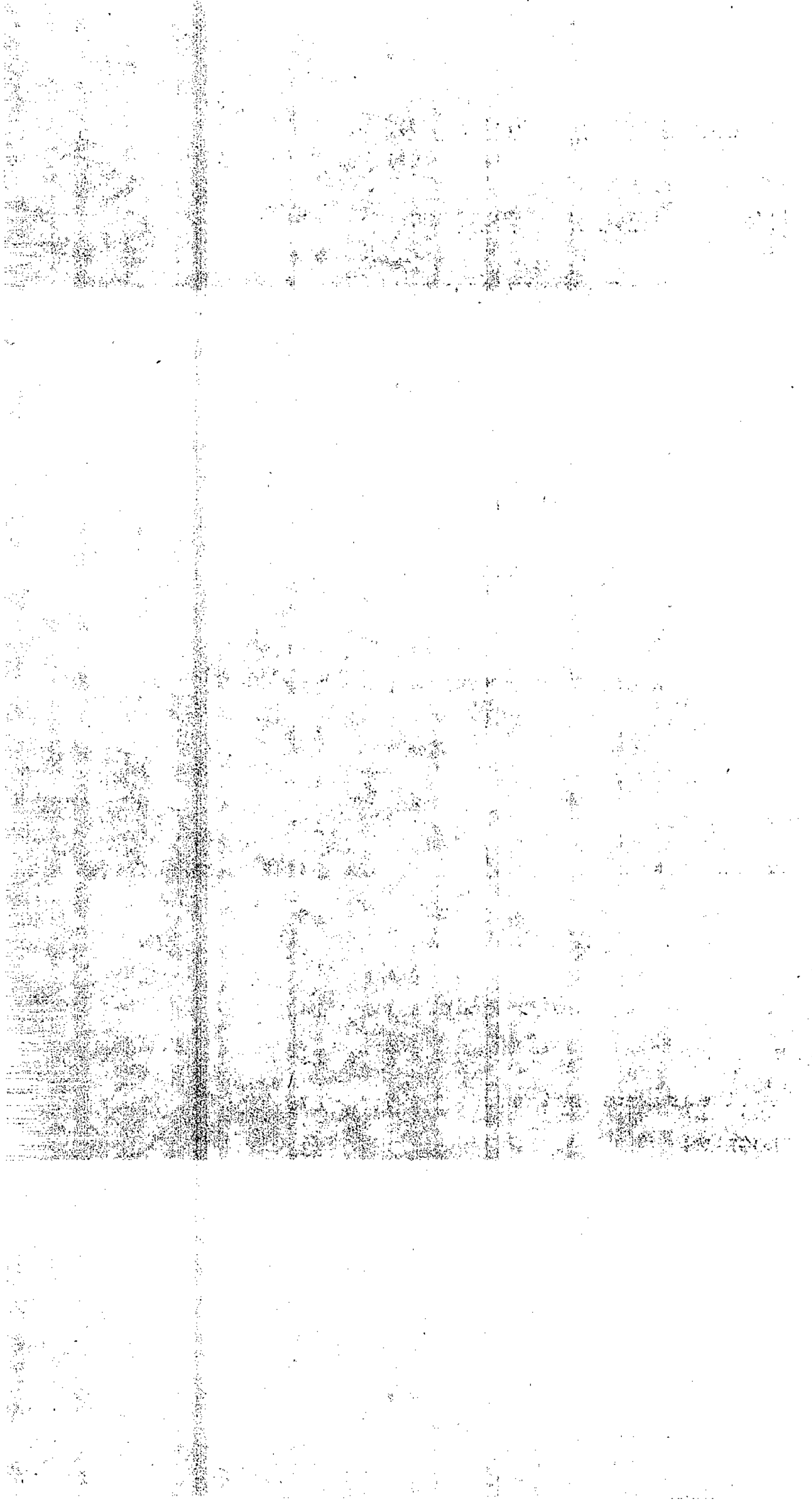


TABLE OF CONTENTS (Continued)

LIST OF FIGURES

INTRODUCTION

INTRODUCTION

The procedures outlined in this manual include both parametric statistics, where the distribution of the data is a known function, and nonparametric statistics, where the distribution of the data is unknown. Several examples of both types are included throughout this manual.

These statistical methods can be applied to all environmental data which can be quantified. Additional nonparametric statistical tests can be used when the data cannot be quantified, but these techniques are not included. The examples contained in this manual all deal with data which pertain to the environmental problem of air pollution, although water pollution examples would have worked equally as well.

It is assumed that the reader has a working knowledge of statistics. Suitable preparation for the material in this manual is offered in the Environmental Protection Agency courses:

801 - Basic Environmental Statistics

804 - Environmental Statistics--Nonparametric

810 - Environmental Statistics--Applied Regression Analysis

Acquiring environmental data is very costly and time consuming. Most of the equipment required for this purpose is very expensive to purchase and operate. Also, to ensure adequate data, sampling must be performed under all types of field conditions in order to cover the range of environmental exposures. Statistical tests described in this manual can be employed in order to:

1. Determine whether the data were collected during a "typical" period based upon historical data (Chi-Squared test for example).

2. Reduce the frequency and/or duration of sampling to minimize costs by:

   a. Determining if the project data are statistically the same as historical data collected at the site which is removed from the project location (Catanova test for example).

   b. Determining whether other data sets collected nearby are statistically similar even though different sampling methods may have been employed (Regression Techniques for example).

   c. Using statistical inference coupled with historical measurements of a related parameter to augment a partial data base (Regression Techniques for example).

   d. Determining whether too much spatial data is being collected (Friedman Test for example).

Much of the data that must be evaluated during the process of environmental impact assessment does not display the characteristics of a "normal" distribution. For example, the distribution of carbon monoxide concentration levels for an entire year at a given site would show a few values at the high concentrations, most of the concentrations at the low concentrations, and no negative values. This is called a right-skewed distribution. If the <u>logs</u> of the concentrations are inserted in place of the arithmetic values, there may be a tendency for the <u>logs</u> of the values to approach a normal distribution.

If the distribution of the data does not appear to follow a normal distribution, and a mathematical data transormation, such as the log transformation discussed above, does not appear to follow a modified normal distribution, then nonparametric statistical tests are employed. Since the distribution of the data is unknown, not as much information is available for the statistical analyses. Therefore, nonparametric statistics are not as powerful as corresponding parametric statistics.

The manual begins with a review of basic concepts of parametric statistics. Included are a discussion of parametric distributions, estimation of central tendency, for both normal and lognormal distributions, and estimation of spread of data and confidence intervals for both normal and lognormal distributions. The equations used to estimate these statistical parameters are included. A brief comparison of parametric and nonparametric statistics is also made.

Part II of this manual contains an example problem which is used throughout the remainder of the chapters to illustrate the application of the various statistical tests presented. Flow charts are shown which can be used to determine when to apply the various tests.

Part III deals with applications of nonparametric statistical tests. Included in this section is a Table indicating when each test should be used and the TENET computer name of each test. Also included are several example problems which describe the use of each test.

The next section, Part IV, deals with the use of regression techniques, a parametric statistical method, which can be used to augment the data base.

Part V describes the Chi-Square Test. The first part describes the chi-square statistic and the rest of the chapter illustrates the use of the statistic to evaluate whether data collected can be assumed to have come from a "typical year".

Part VI deals with methods which can be employed to estimate the worst expected background pollutant levels and typical background pollutant levels which are added to microscale modeling predictions. These values are then compared to ambient air quality standards. Use is made of Larsen's Model to perform this work. Several examples of this type of analysis is included.

Part VII illustrates a method which can be employed to estimate not only the maximum and typical pollutant levels, but also the entire yearly distribution of pollutant levels. This technique makes use of regression equations to augment the data base. An example problem is not included since it would require too much space.

Part VIII deals with the statistical design of air quality surveys. Included are recommendations on sampling techniques, levels of analysis for rural and urban projects, and topics to include when reporting results of statistical tests in technical air quality impact reports.

Part IX contains sample TENET computer runs of each statistical test discussed in the early portions of this manual. Inputs and outputs are included for the user's benefit. The outputs of each computer program are critiqued and the statistical conclusions are discussed.

Several homework problems are included to test the understanding of the material. The manual is concluded with a listing of each computer program so that a user can refer to the listing if the timeshare computer prints an error statement which refers to a line number. Users who do not have access to the TENET system can also add these programs to their system.

BASIC FUNDAMENTALS

INTRODUCTION TO BASIC STATISTICS

The primary purpose in collecting and analyzing aerometric data is to obtain a representative estimate of existing air quality and meteorology within the project area. Air quality can have considerable temporal variation depending on the meteorological conditions and upon the growth of the community. The meteorological conditions that significantly influence air quality are (1) wind speed, (2) wind direction, (3) surface based inversions, and (4) elevated inversions. These meteorological conditions must be considered when collecting and analyzing air quality data. When using the highway line source dispersion model, the concentrations of pollutants are estimated for the most probable as well as for the worst surface meteorological conditions. This involves a particular stability class associated with a prevailing wind speed and direction. To obtain representative air quality data, the samples must be collected under similar meteorological conditions. This will make possible a statistical analysis using data taken from the same population.

The purpose of this section is to review various statistical definitions, concepts, and methods which are necessary for the understanding and usages in analyzing, making inferences, and the presentation of pollutant data. Detailed applications of these basic concepts are illustrated in the remainder of this manual.

GRAPHICAL OR TABULAR PRESENTATION OF DATA

1. Frequency distribution. A frequency distribution is a table which lists classes of data and frequency with which the data in the classes appear.

2. Histogram. A histogram is a graph that represents the class frequencies in a frequency distribution by verticle rectangles.

3. Frequency curve. A frequency curve is a smooth graph derived by plotting the class against its frequency and joining each frequency point by a smooth curve.

4. Cumulative frequency distribution. Cumulative frequency distributions are constructed on either a more than or less than basis. A graphical representation of this distribution is called cumulative frequency curve. These two concepts can best be explained through examples.

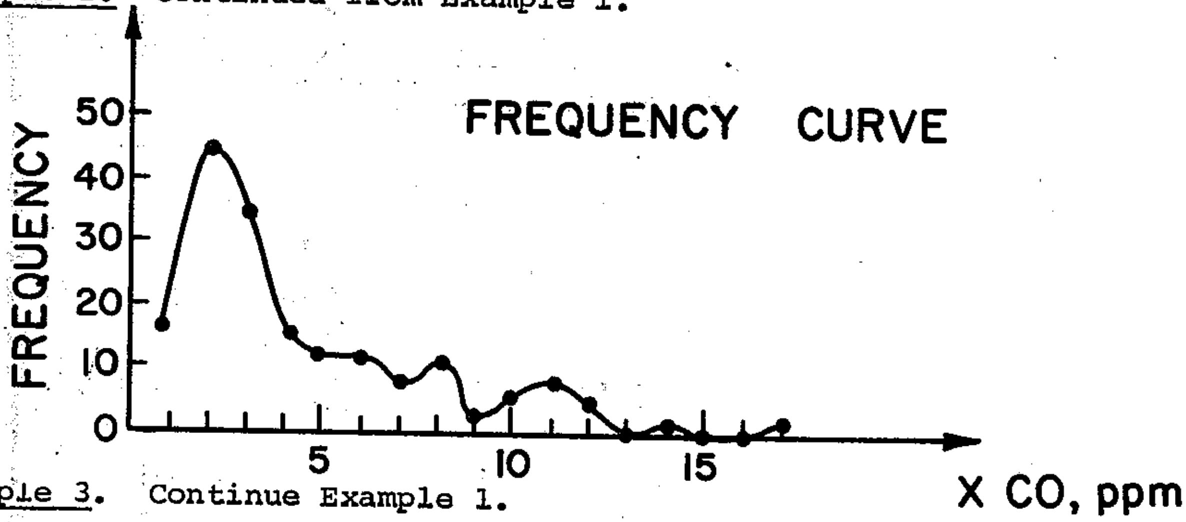
Example 1. Carbon monoxide (CO) pollutant frequency data obtained from BAAPCD Air Monitor Station, San Jose, California.

FREQUENCY DISTRIBUTION

CO hourly averages (ppm)	Frequency From: Alma Street December 1970, 1971 Hours: 0700-0900
17*	1
16	0
15	0
14	1
13	0
12	4
11	7
10	4
9	1
8	11
7	8
6	12
5	12
4	15
3	34
2	45
1	16

*Note: ranking is performed from high to low values for later use.

Example 2. Continued from Example 1.



Example 3. Continue Example 1.

CUMULATIVE FREQUENCY DISTRIBUTION

Concentration Class (i)	Frequency (f_i)	Cumulative Frequency,
17	1	1
14	1	2
12	4	6
11	7	13
10	4	17
9	1	18
8	11	29
7	8	37
6	12	49
5	12	61
4	15	76
3	34	110
2	45	155
1	16	171
	n = 171	

NUMERICAL PRESENTATION

1. Central Tendency

a. Arithmetic mean (or mean) (m)

Definition:

$$m = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where $X_1, X_2, \ldots, X_n$ denote a set of data with n observations (or measurements).

Compute the mean from a frequency distribution:

$$m = \sum_{i=1}^{k} \frac{f_i\, x_i}{\sum_{i=1}^{k} f_i}$$

Where f_i = frequency in class i = 1, 2,...., k classes.

Example 4. Use data from Example 1.

$$m = \frac{1(16) + 2(45) + \cdots + 17(1)}{171} = 4.38$$

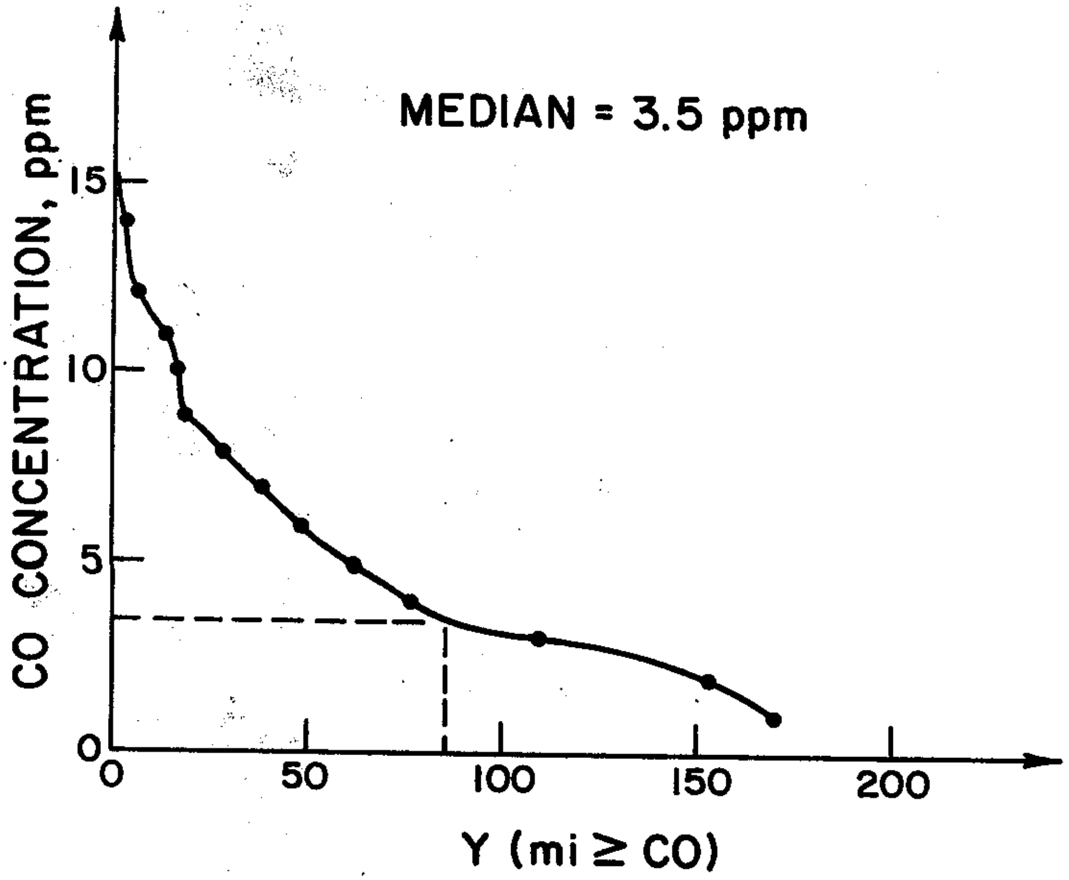
b. Median. The median can be obtained from a cumulative curve. The median is obtained by:

(1) Locate or compute the 50% point on the horizontal scale (Y-axis).

(2) Draw a perpendicular line from the 50% point to intersect the cumulative curve.

(3) At the intersection, drop a perpendicular to the vertical scale (X-axis).

Example. Continue from Example 3



c. Geometric mean (m_g)

Definitions: (1) For a set of nonnegative data, $x_1, x_2, \ldots, x_n$,

$$m_g = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}$$

in logarithmic form:

$$m_g = e^{\left[\frac{1}{n}\sum_{i=1}^{n} \ln x_i\right]}$$

(2) From a frequency table,

$$m_g = e^{\frac{1}{n}\sum_{i=1}^{k} f_i \ln x_i}$$

Where

$$n = \sum_{i=1}^{k} f_i$$

Example 6. Use Example 1 data.

$$m_g = e^{\frac{1}{171}\left[16 \cdot \ln 1 + 45 \ln 2 \ldots + 1 \cdot \ln 17\right]} = e^{1.24} = 3.45$$

2. Dispersion

a. Range. The range is the difference between highest and the lowest measurement.

b. Standard deviation (s).

Definition: (1) For a sample of n measurements, X_1, X_2,...,Xn,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(2) From a frequency table,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2} = \sqrt{\frac{n\sum_{i=1}^{k} f_i x_i^2 - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n(n-1)}}$$

Where $n = \sum_{i=1}^{k} f_i$

Example 7. Use Example 1 data.

$$S = \sqrt{\frac{1}{170}\left[16(1-4.38)^2 + \ldots + 1(17-4.38)^2\right]} = 3.14$$

c. Standard geometric deviation (S_g).

$$S_g = e^{\sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(\ln x_i - \ln m_g)^2 f_i}}$$

Where $n = \sum_{i=1}^{k} f_i$

Example 8. Use Example 1 data.

$$S_g = e^{\sqrt{\frac{1}{170}\left[16(\ln 1-\ln 1.24)^2+\ldots+(\ln 17-\ln 1.24)^2\right]}} = 2.00$$

3. Confidence Intervals

a. Confidence interval for the mean of a normally distributed population is:

(i) for large sample (n > 30),

$$\left(\bar{x}-Z_q*\frac{s}{\sqrt{n}},\ \bar{x}+Z_q*\frac{s}{\sqrt{n}}\right)$$

Where $\bar{x}$ = sample mean

s = sample standard deviation

q = the confidence level

Z_q = critical value, and obtained from the normal probability table (Table 1 on Page 285) with the desired confidence level given.

The following table gives various values of Z_q corresponding to various confidence levels used in practice.

Confidence level (q%)	90	95	99
Z_q	1.645	1.96	2.58

For example, a 95% confidence interval containing the true mean of a pollutant in a specified site is:

$$\left(\bar{x}-1.96\frac{s}{\sqrt{n}},\ \bar{x}+1.96\frac{s}{\sqrt{n}}\right)$$

(ii) For small sample ($n \leq 30$)

$$\left(\bar{x} - t_{(1+q)/2,n-1} \cdot \frac{s}{\sqrt{n}} , \bar{x} + t_{(1+q)/2,n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

Where $t_{q,n-1}$ = critical value corresponding to the confidence level q that is evaluated for n-1 degrees of freedom. It is obtained from the t- probability table (Table 2 on Page 286).

For example, a 95% confidence interval with 21 pollutant measurements is:

$$\left(\bar{x} - 2.086 \frac{s}{\sqrt{n}} , \bar{x} + 2.086 \frac{s}{\sqrt{n}} \right)$$

b. Confidence interval for the mean of a lognormal distributed population is ($n > 100$):

$$\left(m_g \cdot S_g^{-Z_q/\sqrt{n}} , m_g \cdot S_g^{Z_q/\sqrt{n}} \right)$$

Where M_g = geometric mean

S_g = standard geometric deviation

Z_q = defined above

n = sample size

G. CONFIDENCE INTERVALS

A NOMOGRAPHIC REPRESENTATION FOR THE CONFIDENCE INTERVAL AND LEVEL OF SIGNIFICANCE IS PRESENTED BELOW.

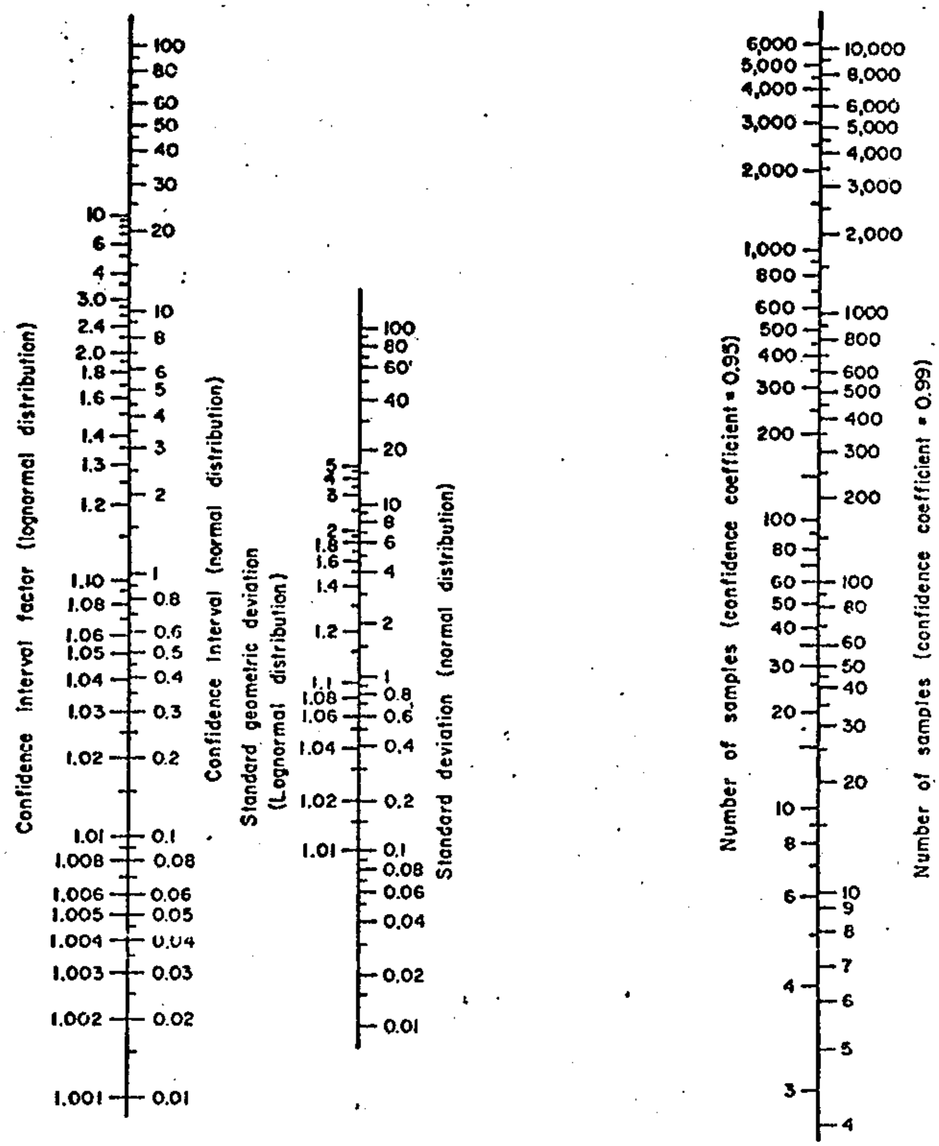


Figure 1. Confidence intervals for the mean of a group of random samples.
For a normal distribution, use numbers on right sides of center and left scales. A straight line connects related values of the confidence interval (above and below the mean), the standard deviation, and the number of samples (n) in the group for either 0.95 or 0.99 confidence coefficients (calculated from the t_{n-1} distributions). If desired, the numbers on both the standard deviation scale and on the confidence interval scale may be multiplied by the same factor (e.g. 0.1, 0.01, 0.001, etc.)
For a lognormal distribution, use numbers on left sides of center and left scales. A straight line connects related values of the confidence interval factor (with which to multiply or divide the geometric mean), the standard geometric deviation, and the same scales for number of samples in the group.

SOURCE: "Simplified Methods for Statistical Interpretation of Monitoring Data" by B.E. Saltzman Journal of Air Pollution Control Association, Feb. 1972.

For example, a 95% confidence interval containing the true geometric mean for Example 1 data on Page 9 is

$$(3.45 \times 2^{-.15} \quad , \quad 3.45 \times 2^{+.15}) = (3.11 , 3.83$$

A graphical method of determining these values is presented in Figure 4. The procedure is to divide the geometric mean by the confidence interval factor for the lower value and multiply the geometric mean by the confidence interval factor for the upper value:

Lower Value	Upper Value
$\frac{3.45}{1.11} = 3.11$	$3.45 \times 1.11 = 3.83$

NORMAL AND LOGNORMAL DISTRIBUTIONS

Various theoretical distributions can be used to describe pollutant concentrations. The most common and appropriate ones are the normal and lognormal distributions. The normal distribution is characterized by two parameters - the mean and the standard deviation.

1. The equation of the normal distribution curve is:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Where π = 3.1416, a constant

e = 2.7183, a constant

σ = the standard deviation, a parameter

μ = the mean, a parameter

X = abscissa, measurement on the horizontal axis

Y = ordinate, height of curve corresponding to a value of X.

2. Fitting a normal curve to a frequency table, the equation is:

$$Y = \frac{n}{S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-m}{s}\right)^2}$$

Where n = total number of samples
m, s = sampled mean & standard deviation respectively.

Example 9. Given the following set of pollutant data, fit a normal curve to it.

Concentration in ppm	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	2	3	4	4	5	10	15	12	8	7	6	4	4	1

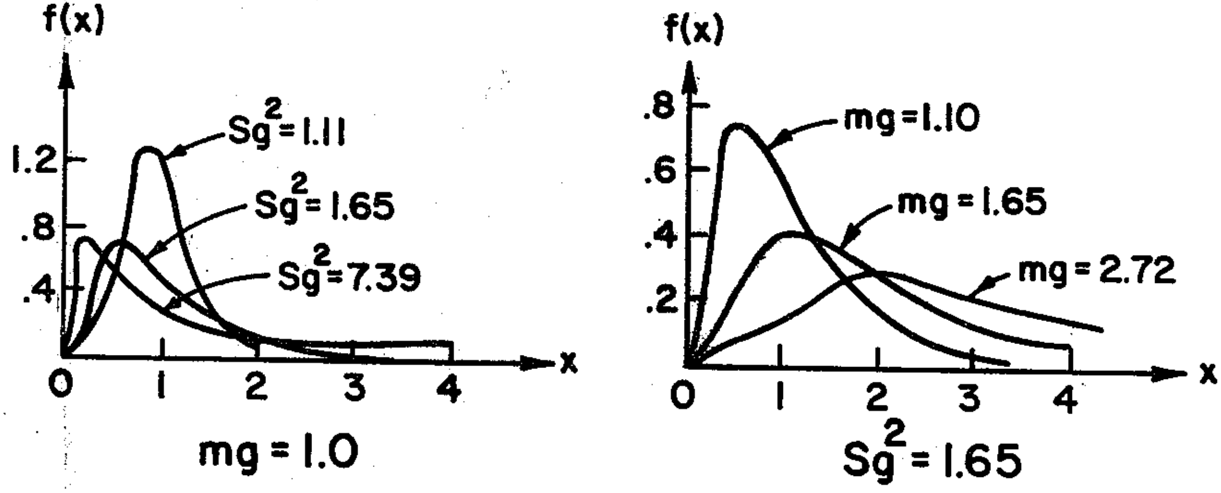
LOGNORMAL DISTRIBUTIONS

1. The lognormal distribution is a positive skewed frequency curve which has the same shape as a normal curve when the concentrations are plotted on a logarithmic scale. That is, the logarithms of the concentrations are normally distributed. The distribution is:

$$f(x) = \frac{1}{X \cdot \ln S_g \sqrt{2\pi}} e^{-\frac{1}{2\ln S_g^2}\left(\ln \frac{X}{m_g}\right)^2}, \quad X > 0$$

Where m_g = the geometric mean
s_q = the standard geometric deviation
$f(x)$ = the amount of concentrations in ppm at each level of X.

Example 10.



For a lognormal distribution, the arithmetic mean (m), geometric mean (m_g), standard deviation (s), and standard geometric deviation (S_g) are related as follows:

$$S_g = e^{(\ln(s^2/m^2+1))^{.5}}$$

$$m_g = m \times e^{-\frac{1}{2}(\ln S_g)^2}$$

LOGNORMAL PROBABILITY PAPER AND PLOTS

The objective of this section is to show how to plot a set of air pollutant concentration data in a lognormal probability paper* for the following two usages:

*For further usages for analyzing air pollutant data, see applications of Larsen's model for air quality analysis (Reference 5).

(1) to test the assumed lognormal distribution of concentration data**, and

(2) to estimate the lognormal distribution parameters (m_g and s_g) directly without the need for tedious calculations if (1) is true.

To test the assumed lognormal distributions, the steps are as follows:

(a) construct a modified relative percentage cumulative frequency distribution by 100 $m_i/(n+1)$,

where m_i = cumulative frequency of measurements in class i, ($\geq$pollutant concentration).

n = total frequency, total number of measurements in a given set.

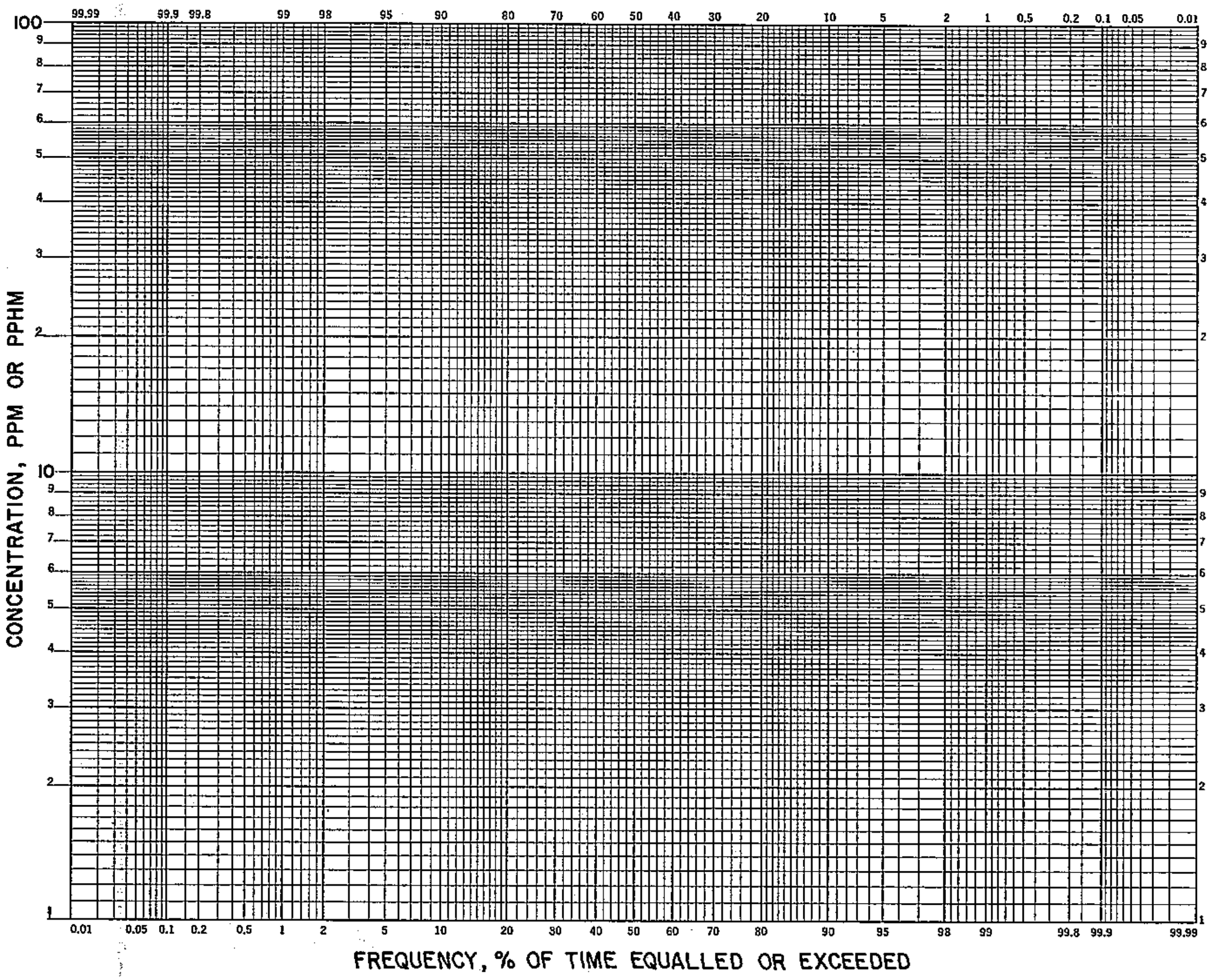
(b) plot the relative cumulative values on a lognormal probability paper (Figure 5 is an example of lognormal probability paper). Concentrations (X_i) are plotted on a log ordinate and the relative cumulative values (Y_i) are plotted on the abscissa.

(c) if a straight line gives a close fit to the plotted points, then the assumed distribution is correct. If the result is a curved plot, it indicates that the assumed distribution is incorrect.

**In testing the goodness-of-fit to an assumed distribution, the Chi-square or the Kolmogorov-Smirnov test can be used if a precise statistical test is needed.

Figure 5

LOG-NORMAL PROBABILITY PAPER



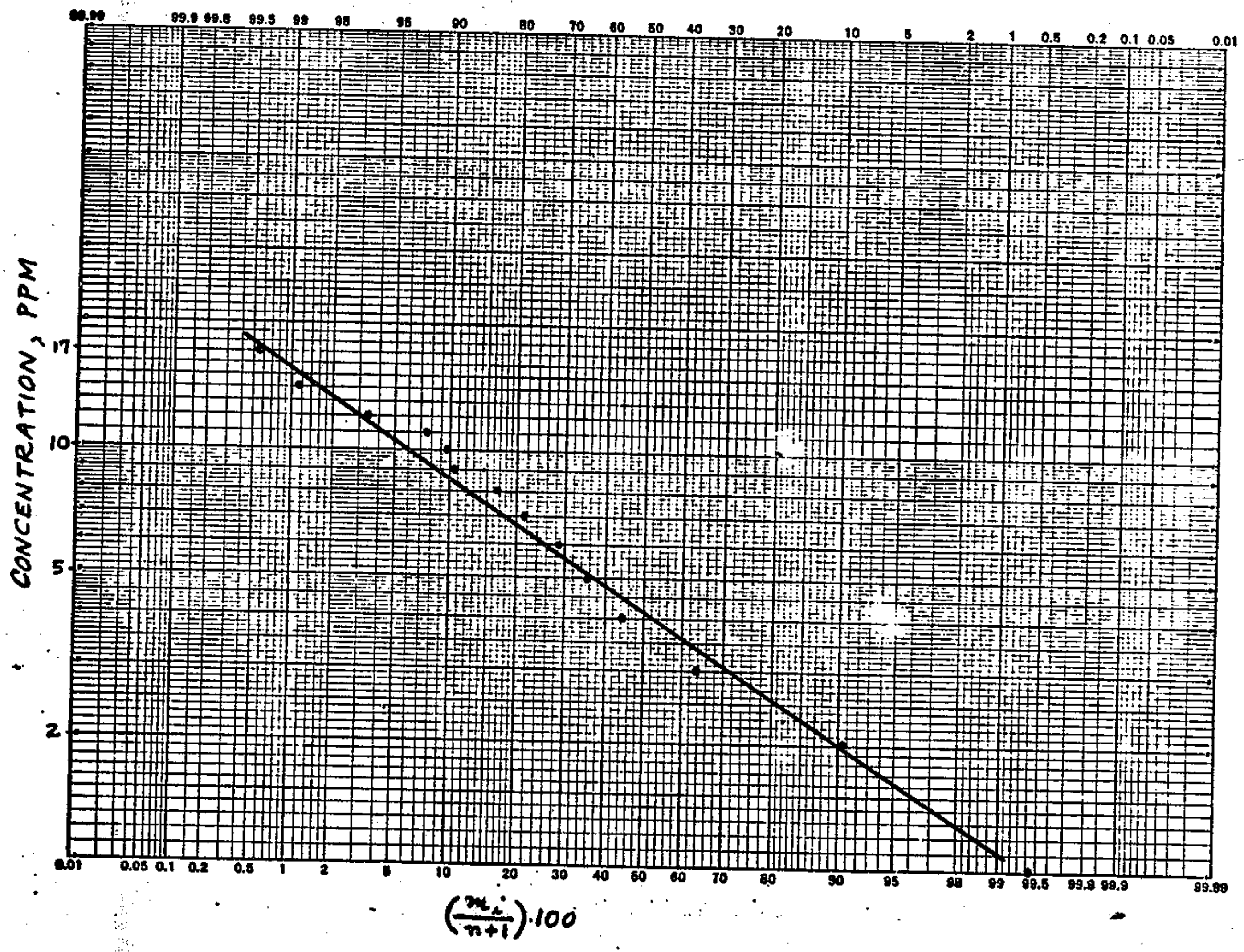
Example 11. Test whether the data obtained in Example 1 is log normally distributed.

Solution:

(a)

CO Concentration	Cumulative Frequency, $m_i \geqslant co$	
17	1	(1/172)(100)= .58
14	2	(2/172)(100)= 1.16
12	6	" 3.49
11	13	" 7.56
10	17	" 9.88
9	18	" 10.47
8	29	" 16.86
7	37	" 21.51
6	49	" 28.49
5	61	" 35.47
4	76	" 44.19
3	110	" 63.95
2	155	" 90.17
1	171	" 99.42

(b) PLOTTING THE CONCENTRATION VALUES VS. THE RELATIVE CUMULATIVE VALUES ON A LOG-NORMAL PAPER :



(c) A straight line gives a good approximation to the plotted points; therefore, this indicates that we can assume this set of data is log-normally distributed.

Estimates of m_g and S_g can be made directly from the plot when the lognormal assumption is true by using the following relationships:

$m_g = X_{50}$, the median, the concentration at the 50% frequency point on the line.

$S_g = \frac{X_{16,}}{X_{50}}$ the ratio of the 16 percentile to the 50 percentile on the line.

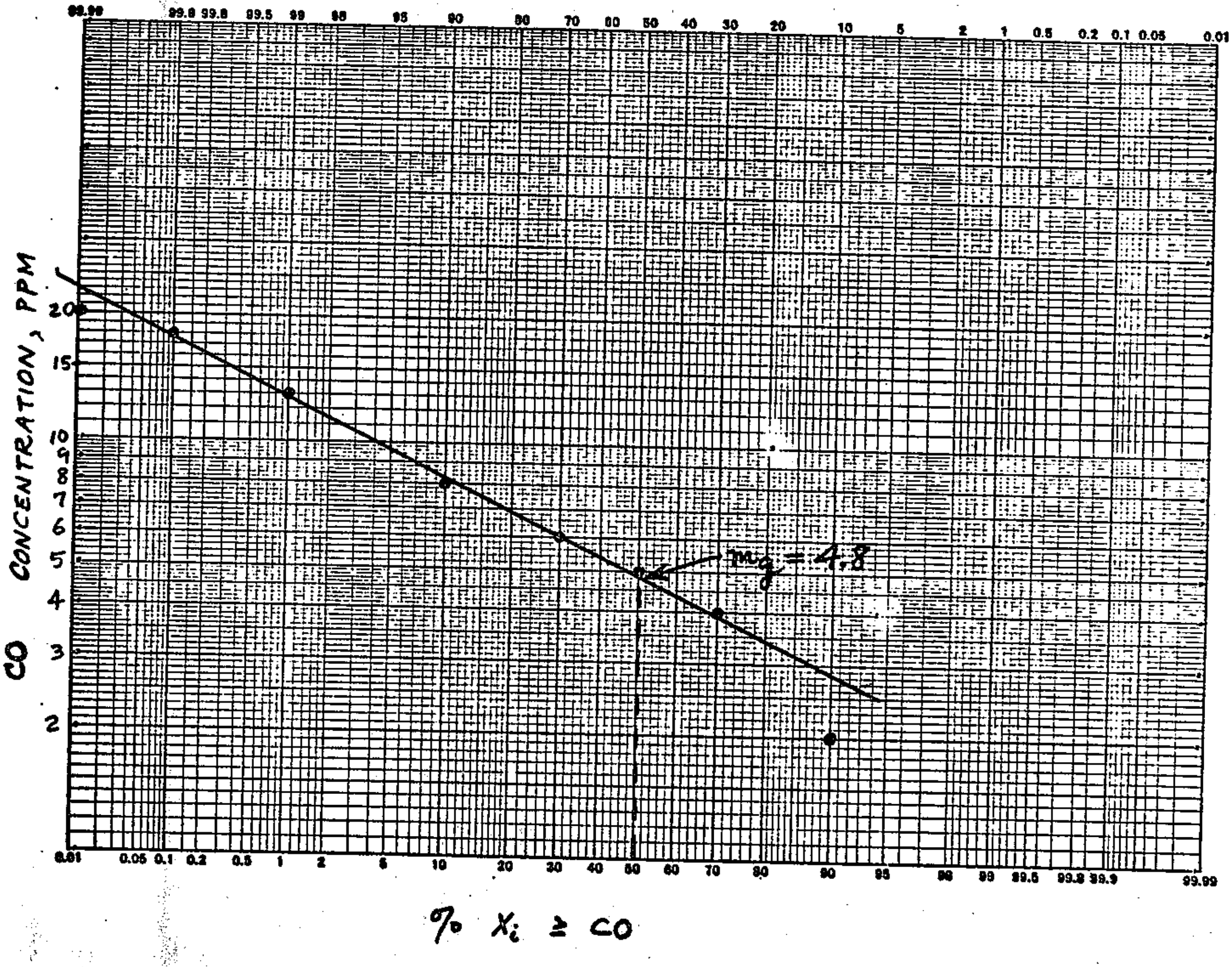
Example 12. To test whether or not the data - one-hour average CO concentrations at San Francisco in the period of 1962 through 1968 - is lognormally distributed. If it is, estimate the parameters, m_g and S_g.

Solution:

Data obtained by Larsen:

CO, ppm (X)	Relative Cumulative Frequency (%$X_i \geq$ CO)
20	.01
18	.10
13	1.00
8	10.00
6	30.00
5	50.00
4	70.00
2	90.00

PLOTTING THE ABOVE SET OF DATA ON A LOG-NORMAL PROBABILITY PAPER :



Notice a straight line fits the plotted points well; therefore, the data is log-normally distributed.

Also, $m_g = X_{50} = 4.8$ ppm

$$s_g = \frac{X_{16}}{X_{50}} = \frac{7.2}{4.8} = 1.52 \text{ ppm}$$

HYPOTHESIS TESTING

Hypothesis testing is the process of inferring from a sample whether or not to accept a certain assumption about the population. The assumption itself is called the hypotheses.

Examples:

1. Site 1 has higher overall mean L_{50} noise level than Site 2.

2. The CO concentration at Site A is higher than at Site D for the summer season.

3. The suspended sediment loads in the upper part is lower than in the lower part of Stream X.

In each case, the hypotheses is tested on the basis of the evidence contained in a random sample, and the final decision is either to reject or to accept the hypothesis.

The 9-step general procedure for the test of a hypothesis is as follows:

1. State the experimental goal.

2. State the statistical null hypothesis (H_o) and the alternative (H_1). H_o there is no statistical significant difference, H_1 there is a statistical difference.

3. Choose the level of significance, α, the chance of rejecting the hypothesis if it is true (the choice is set arbitrarily but $\alpha = 0.05$ is generally recommended).

4. Choose an appropriate statistical test for testing the null Hypothesis (Objectives of this Course).

5. Find (or assume) the sampling distribution of the statistical test under the assumption that the null hypothesis is true.

6. Define the region of rejection (critical region), those values of the statistical test which would cause the rejection of the null hypothesis.

7. Compute the value of the statistical test from the sample and see whether or not it falls in the rejection region.

8. State the statistical conclusion, a statement of acceptance or rejection of the null hypothesis. If the value of the statistical test falls into the rejection region, the null hypothesis is rejected; if not, accept the null hypothesis.

9. State the experiemental conclusion in light of the acceptance or rejection of the null hypothesis.

A significance level (α), is a level which indicates the probability that the null hypothesis is correct or not correct. Usually this is taken to be .05 which indicates that the probability that the null hypothesis is correct exceeds 1.0 -.05 or .95. This means that 95 times out of 100, we made the right decision.

Significant is used in the statistical sense of the word and means that the probability of the observed difference being due to chance alone is equal to the level of significance.

Note: A difference may be statistically significant but not physically important and therefore, not significant from a practical point of view.

DEGREES OF FREEDOM

Degrees of freedom (denoted by ν or df, or D.F.) are equal to the number of independent observations in the sample minus the number of population constraints which must be estimated from sample observations.

For example:

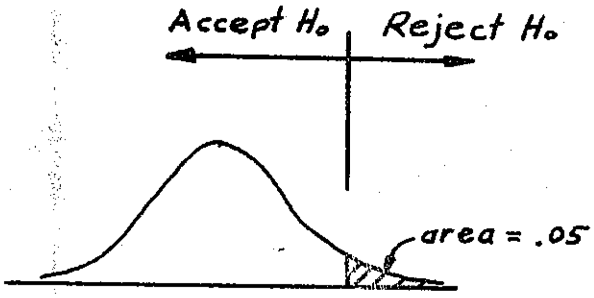
(i) For the t- statistic, $\nu = n-1$, where we must estimate μ, the population parameter.

(ii) For the χ^2-statistic, $\nu = n-1$, the one df subtracted was for estimating the population parameter σ^2.

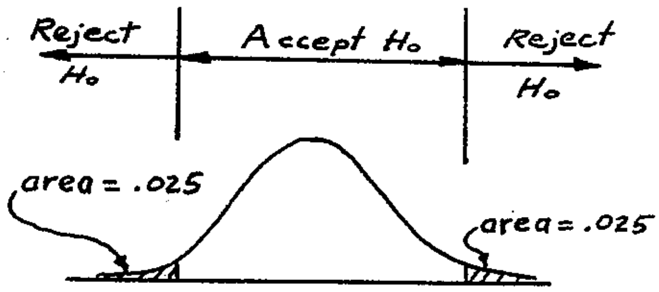
CRITICAL REGION

The rejection region is a region of the sampling distribution. It is a set of all points in the sample space which result in the decision to reject the null hypothesis. The set of all points in the sample space not in the rejection region is called the acceptance region. The location of the rejection region is affected by the nature of H_1. If H_1 indicates the predicted direction of the difference, then a one-tailed test is called for; otherwise a two-tailed test is called for. One-tailed and two-tailed tests differ in the location (not in the size) of the rejection region. A one-tailed test has the rejection region at one end (or tail) of the sampling distribution. In a two-tailed test, the rejection is located at both ends of the sampling distribution.

The size of the rejection region is expressed by α, the level of significance. For example, if $\alpha = .05$, then the size of the rejection region is 5% of the area under the curve in the sampling distribution. One-tailed and two-tailed regions of rejection for $\alpha = .05$ are illustrated in the following figure. Observe that these two regions differ in location but not in total size (area).



A one-tailed region of rejection when $\alpha = .05$



A two-tailed region of rejection when $\alpha = .05$

PARAMETRIC VERSUS NONPARAMETRIC STATISTICS

I. Parametric Statistics

1. Observations must be independent.

2. Observations must be drawn from normally distributed population.

3. These populations must have the same variance.

II. Nonparametric Statistics

1. Data is continuous.

2. No assumptions made about distribution of data.

3. Observations are made independently of each other.

4. Most appropriate for non-random data.

5. Can be used for small sample size.

Disadvantage of Nonparametric Statistics:

1. Do not usually use original data for inference.

Devices Used in Nonparametric Tests

1. Counting of categories (+ and - signs).

2. Arrangement of observations in order of magnitude.

3. Assignment of ranks to such an order.

4. Use of the median instead of the mean.

5. Use of the quartile deviation instead of the standard deviation.

Uses of Nonparametric Statistics

1. Design of air quality surveys.

2. Correlation studies - meteorological wind roses.

3. Presentation of data.

NONPARAMETRIC TESTING

Nonparametric methods are considered "distribution free" since they make no assumption with regard to the distribution of measurements in the population, whereas parametric methods, such as the t tests and F tests assume normally, or lognormally distributed measurements. Two basic assumptions are associated with most nonparametric statistical tests: the measurements are independent and the variable under study has underlying continuity. When we wish to test whether or not two samples are from the same population using nonparametric tests, we are generally interested in the "location" difference. One method of discovering a possible difference in location between the two sampling distribution curves is by use of the medians. Another method is by the predominance of higher values in one sample than in the other. Most nonparametric tests are based on ranks, signs, or groups.

Wilcoxon Test

1. We wish to know if the two related samples come from the same population.

2. H_o: Two samples are identical with respect to their locations.

   H_1: Two samples are different.

3. Choose $\alpha = 0.05$.

4. Calculate $d_i = X_i - Y_i$, the signed difference between two measurements X_i , Y_i

5. Rank these d_i without respect to sign. When ties occurred, assign the average of the tied ranks.

6. Affix to each rank the sign (+ or -) of the which it represents.

7. Use statistic W = the smaller of the sums of the like-signed ranks.

8. Determine n = the total number of d's having a sign.

9. For $n \leq 20$, rejection regions are $W < W_{\alpha/2}$ and $W > W_{\alpha/2}$ from Table 5 for a given significance level. Reject H_o if W falls into the rejection region. For $n > 20$ use normal approximation.

10. State the experimental conclusion.

Example 13

CO concentrations are measured at two sites for six different days in the morning period. The data are recorded as follows:

Hour	Site 1						Site 2					
0600-0700	7.1	8.2	6.0	7.1	7.9	8.1	7.5	8.0	6.2	7.0	7.8	6.0
0700-0300	6.9	7.8	9.2	9.4	9.4	9.8	7.5	7.7	8.1	7.6	7.8	9.0
0800-0900	5.1	5.2	5.8	5.6	6.2	6.4	7.0	7.1	6.8	6.4	6.0	6.6

Analyze the problem with a Wilcoxon signed-rank test.

Solution:

Steps:

(2) H_o: No difference in CO concentrations between Sites 1 and 2.

H_1: There is some difference.

(3) Choose $\alpha = .05$.

(4)-(6):

	Site				
Hour	1 (X)	2 (Y)	d = X-Y	Rank of $\lvert d \rvert$	Rank with less Frequent sign
1	7.1	7.5	- .4	-8	8
	8.2	8.0	+ .2	4.5	
	6.0	6.2	- .2	-4.5	4.5
	7.1	7.0	.1	1.5	
	8.1	7.8	.3	7	
	7.9	6.0	1.9	17	
2	6.9	7.5	- .6	-9	9
	7.8	7.7	.1	1.5	
	9.2	8.1	1.1	13	
	9.4	7.6	1.8	15	
	9.4	7.8	1.6	14	
	9.8	9.0	.8	10.5	
3	5.1	7.0	-1.9	-17	17
	5.2	7.1	-1.9	-17	17
	5.8	6.8	-1.0	-12	12
	5.6	6.4	- .8	-10.5	10.5
	6.2	6.0	.2	4.5	
	6.4	6.6	- .2	- 4.5	4.5

$\Sigma = W = \underline{82.5}$

(7) n = 18

(8) $W_{.025} = 41$)
$W_{.975} = 130$) From Table 5; Accept H_o.

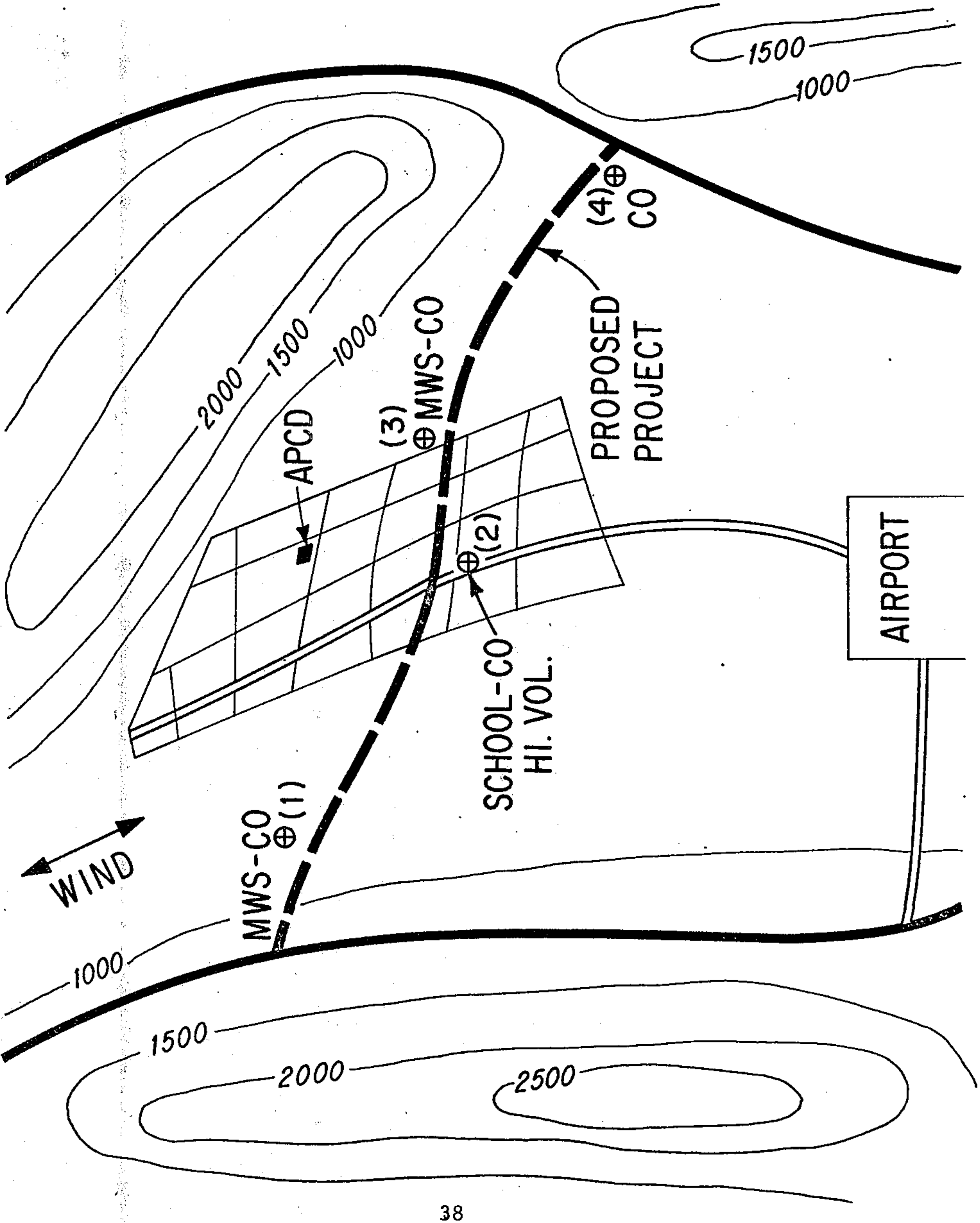
(9) From the evidence of the samples, there is no reason to reject the hypothesis that the sites do not have the same CO distribution at the 5% significance level. Other nonparametric testing procedures can be obtained in References [7] and [8].

FLOW CHARTS FOR STATISTICAL ANALYSES

The general problem presented as Figure 1 will be used later in this manual to demonstrate all the statistical techniques presented. The steps required to perform a meteorological survey are outlined in Figure 3 "Flow Chart for a Meteorological Survey". The procedures required to perform this work are described on following pages. The steps required to perform an ambient air quality survey are outlined in Figure 4, "Flow Chart for an Ambient Air Quality Survey". The procedures required to analyze the above are also described on the following pages. Refer to Manuals Numbers 1 and 6 for more information regarding meteoroogical or ambient air quality surveys.

The general thought processes for an air quality study that should be followed, as suggested by John Grisinger of District 07, Los Angeles, are outlined in Figure 6. The top line is for a meteorological survey. The steps "correlate meteorological data" and "extrapolate to determine meteorological parameters for project areas" are expanded upon in Figure 7. The bottom line of Figure 6 is for an ambient air quality survey. The steps "correlate air quality data" and "extrapolate to determine air quality parameters for project area" are explained in Figure 8. It should be emphasized that no air quality data should be monitored unless sufficient meteorological data is available to help explain the variations in the data.

Figure 1


1500
1000
(4)⊕ CO
PROPOSED PROJECT
2000
1500
1000
(3) ⊕MWS-CO
APCD
⊕(2)
SCHOOL-CO HI. VOL.
AIRPORT
WIND
MWS-CO ⊕(1)
1000
1500
2000
2500

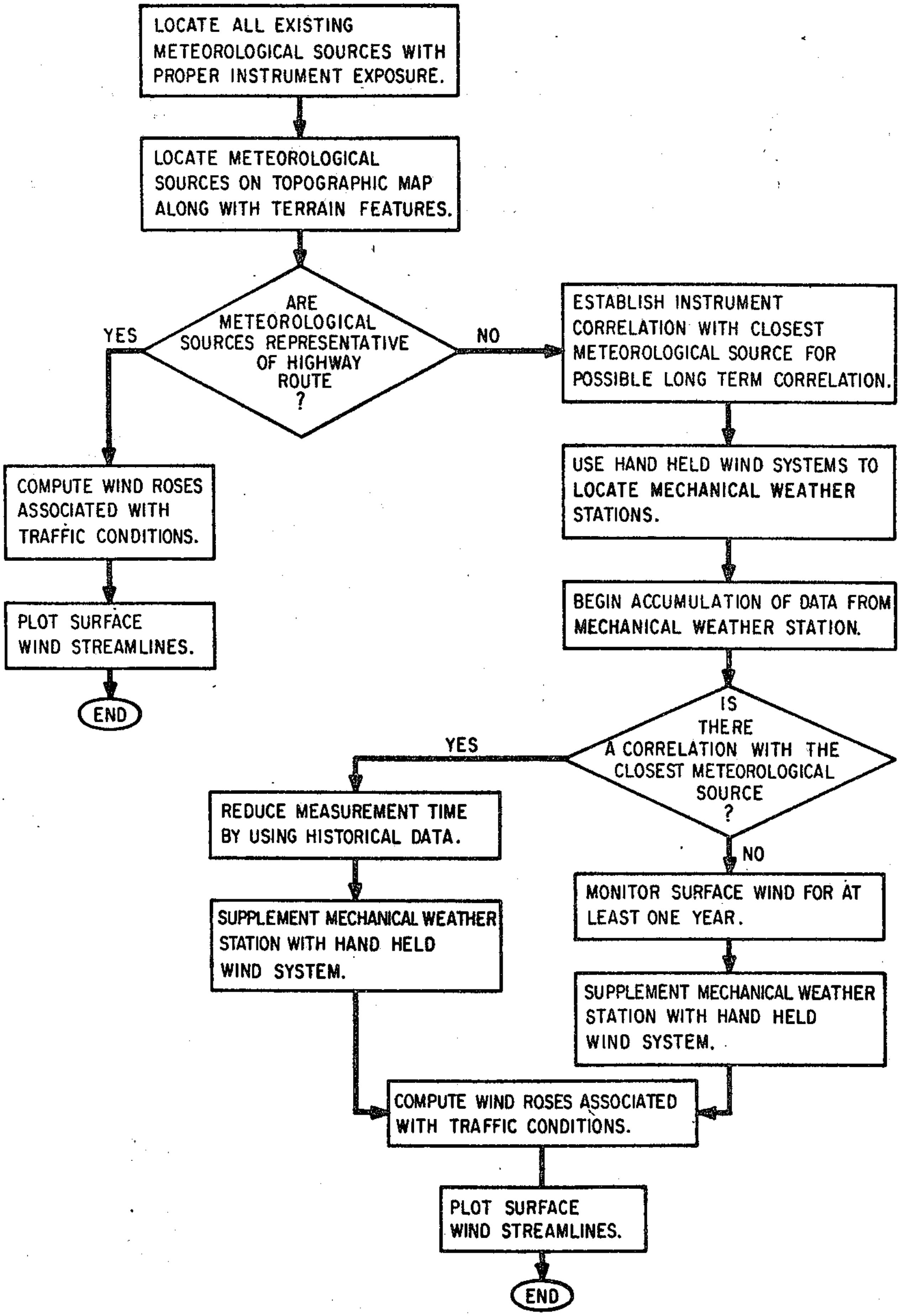

LOCATE ALL EXISTING METEOROLOGICAL SOURCES WITH PROPER INSTRUMENT EXPOSURE.
LOCATE METEOROLOGICAL SOURCES ON TOPOGRAPHIC MAP ALONG WITH TERRAIN FEATURES.
ARE METEOROLOGICAL SOURCES REPRESENTATIVE OF HIGHWAY ROUTE ?
YES
NO
ESTABLISH INSTRUMENT CORRELATION WITH CLOSEST METEOROLOGICAL SOURCE FOR POSSIBLE LONG TERM CORRELATION.
COMPUTE WIND ROSES ASSOCIATED WITH TRAFFIC CONDITIONS.
USE HAND HELD WIND SYSTEMS TO LOCATE MECHANICAL WEATHER STATIONS.
PLOT SURFACE WIND STREAMLINES.
BEGIN ACCUMULATION OF DATA FROM MECHANICAL WEATHER STATION.
END
IS THERE A CORRELATION WITH THE CLOSEST METEOROLOGICAL SOURCE ?
YES
NO
REDUCE MEASUREMENT TIME BY USING HISTORICAL DATA.
MONITOR SURFACE WIND FOR AT LEAST ONE YEAR.
SUPPLEMENT MECHANICAL WEATHER STATION WITH HAND HELD WIND SYSTEM.
SUPPLEMENT MECHANICAL WEATHER STATION WITH HAND HELD WIND SYSTEM.
COMPUTE WIND ROSES ASSOCIATED WITH TRAFFIC CONDITIONS.
PLOT SURFACE WIND STREAMLINES.
END

Figure 3 FLOW CHART FOR A METEOROLOGICAL SURVEY

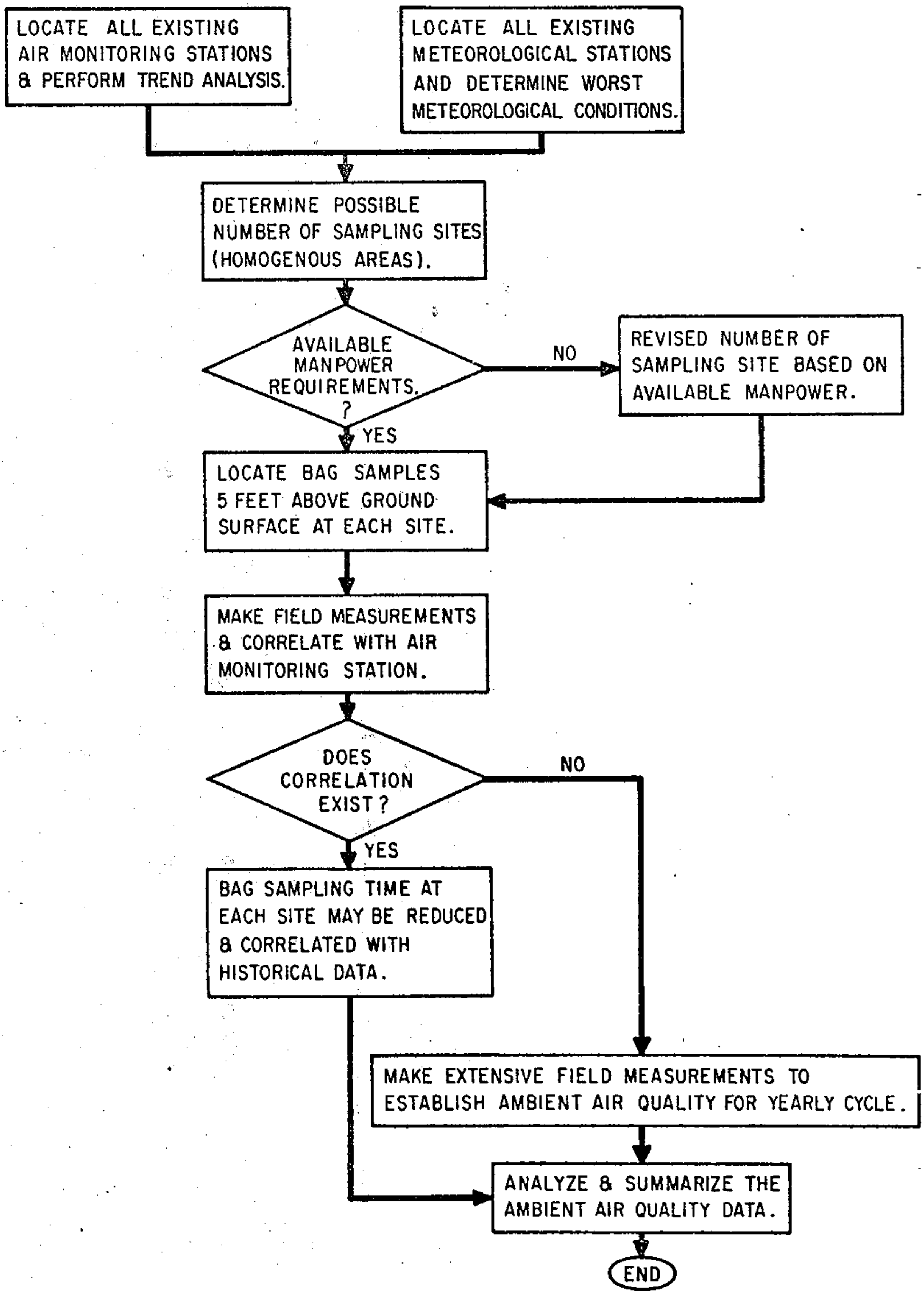
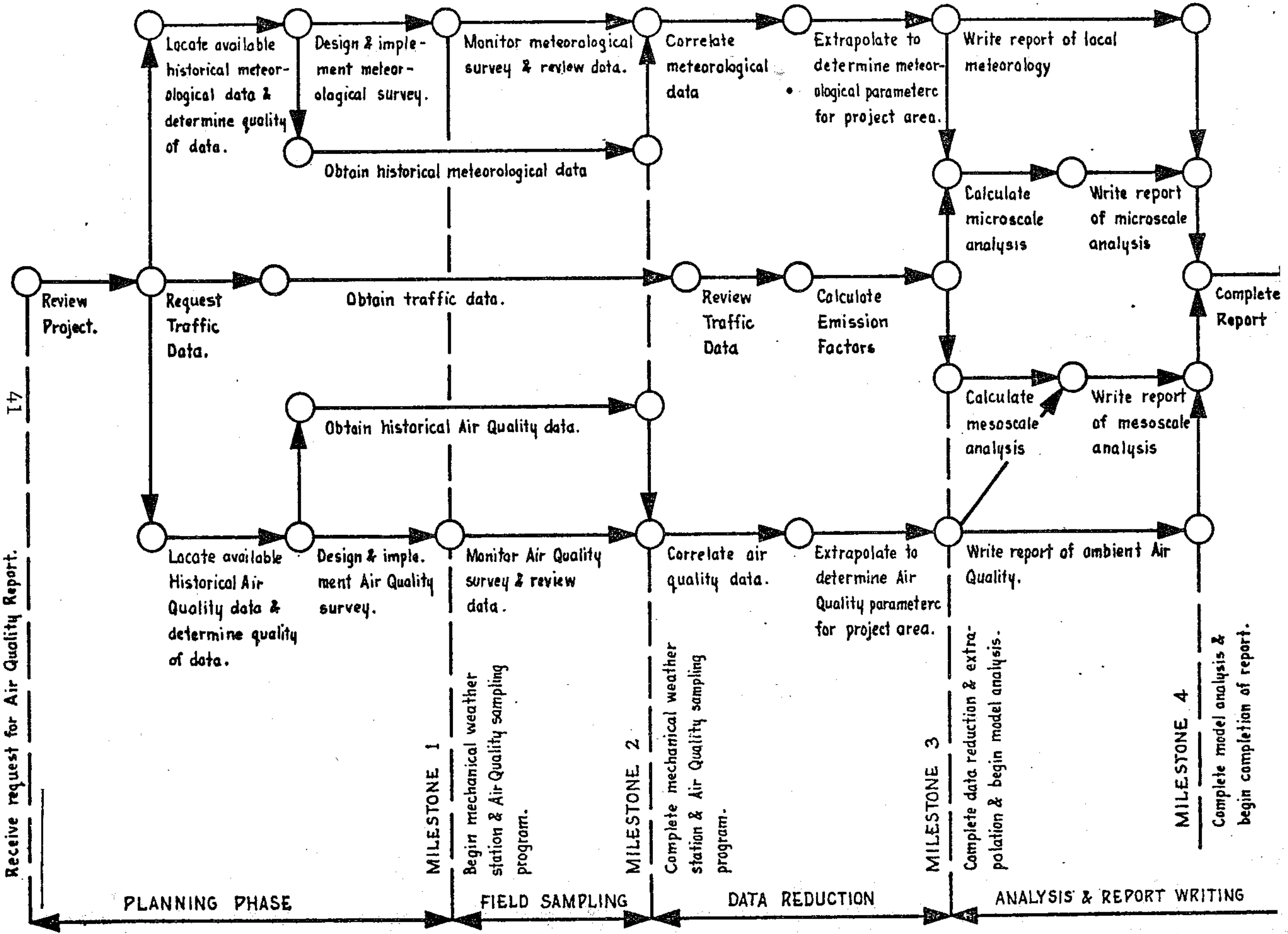
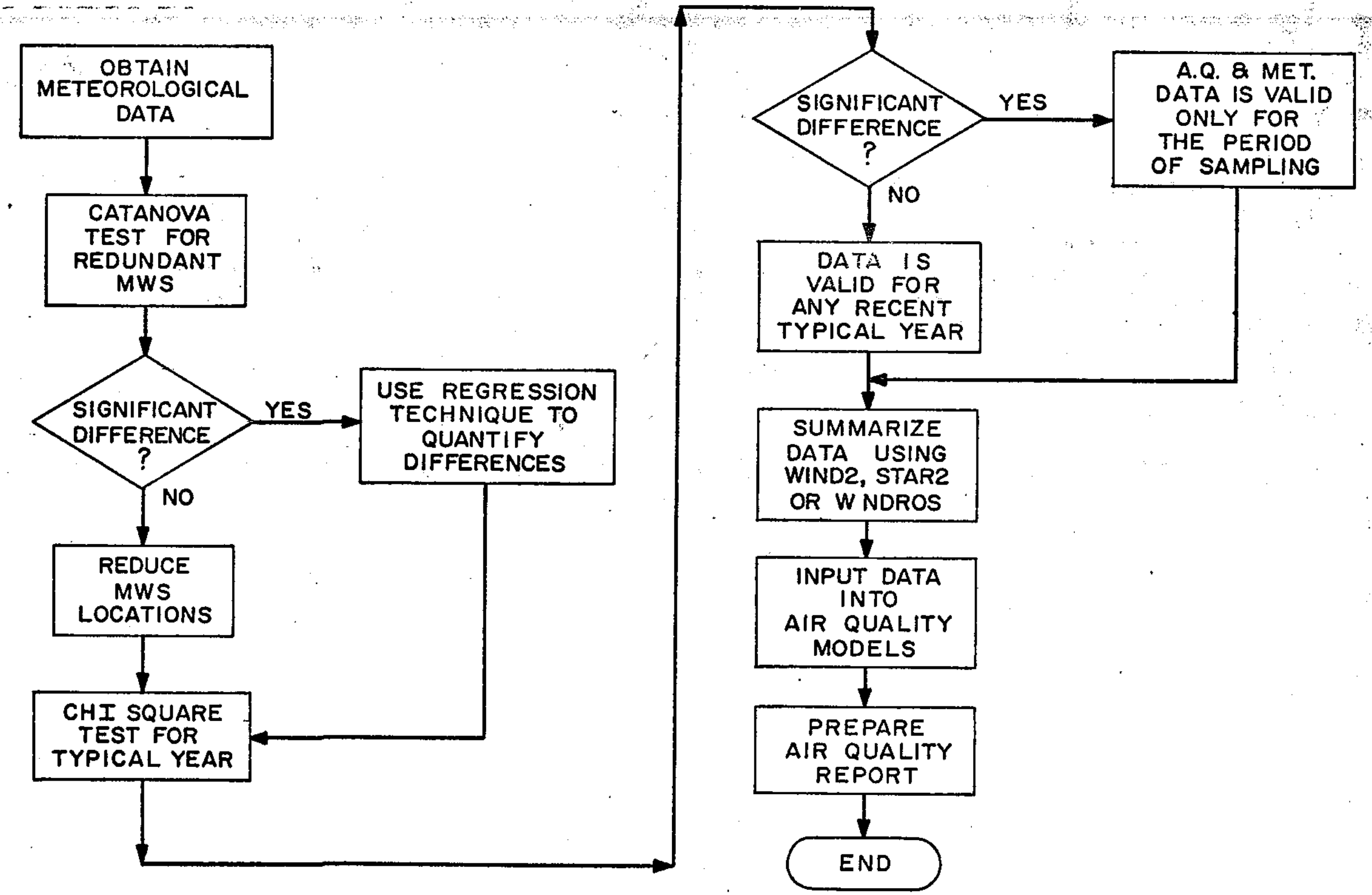

LOCATE ALL EXISTING AIR MONITORING STATIONS & PERFORM TREND ANALYSIS.
LOCATE ALL EXISTING METEOROLOGICAL STATIONS AND DETERMINE WORST METEOROLOGICAL CONDITIONS.
DETERMINE POSSIBLE NUMBER OF SAMPLING SITES (HOMOGENOUS AREAS).
AVAILABLE MANPOWER REQUIREMENTS. ?
NO
REVISED NUMBER OF SAMPLING SITE BASED ON AVAILABLE MANPOWER.
YES
LOCATE BAG SAMPLES 5 FEET ABOVE GROUND SURFACE AT EACH SITE.
MAKE FIELD MEASUREMENTS & CORRELATE WITH AIR MONITORING STATION.
DOES CORRELATION EXIST ?
NO
YES
BAG SAMPLING TIME AT EACH SITE MAY BE REDUCED & CORRELATED WITH HISTORICAL DATA.
MAKE EXTENSIVE FIELD MEASUREMENTS TO ESTABLISH AMBIENT AIR QUALITY FOR YEARLY CYCLE.
ANALYZE & SUMMARIZE THE AMBIENT AIR QUALITY DATA.
END

Figure 4 FLOW CHART FOR AN AMBIENT AIR QUALITY SURVEY

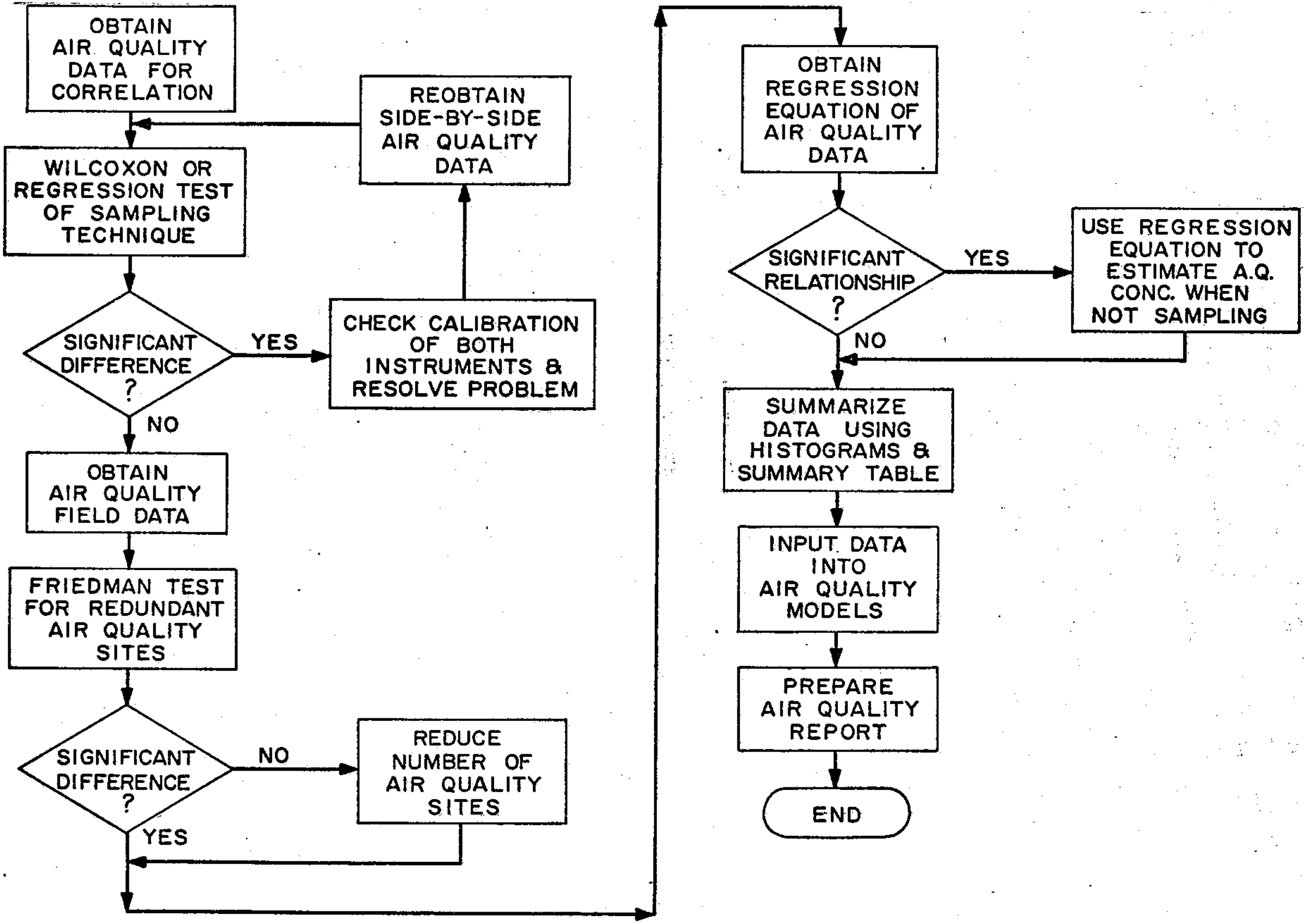
AIR QUALITY STUDY
FOR IMPACT OF HIGHWAYS ON AIR ENVIRONMENT


Review Project.
Request Traffic Data.
Locate available historical meteorological data & determine quality of data.
Design & implement meteorological survey.
Obtain historical meteorological data
Monitor meteorological survey & review data.
Correlate meteorological data
Extrapolate to determine meteorological parameterc for project area.
Write report of local meteorology
Calculate microscale analysis
Write report of microscale analysis
Obtain traffic data.
Review Traffic Data
Calculote Emission Factors
Complete Report
Obtain historical Air Quality data.
Calculate mesoscale analysis
Write report of mesoscale analysis
Locate available Historical Air Quality data & determine quality of data.
Design & imple. ment Air Quality survey.
Monitor Air Quality survey & review data.
Correlate air quality data.
Extrapolate to determine Air Quality parameterc for project area.
Write report of ambient Air Quality.
Receive request for Air Quality Report.
MILESTONE 1
Begin mechanical weather station & Air Quality sampling program.
MILESTONE 2
Complete mechanical weather station & Air Quality sampling program.
MILESTONE 3
Complete data reduction & extrapolation & begin model analysis.
MILESTONE 4
Complete model analysis & begin completion of report.
PLANNING PHASE
FIELD SAMPLING
DATA REDUCTION
ANALYSIS & REPORT WRITING

STATISTICAL ANALYSIS FLOW CHART FOR METEOROLOGICAL SURVEYS


OBTAIN METEOROLOGICAL DATA
CATANOVA TEST FOR REDUNDANT MWS
SIGNIFICANT DIFFERENCE ?
YES
USE REGRESSION TECHNIQUE TO QUANTIFY DIFFERENCES
NO
REDUCE MWS LOCATIONS
CHI SQUARE TEST FOR TYPICAL YEAR
SIGNIFICANT DIFFERENCE ?
YES
A.Q. & MET. DATA IS VALID ONLY FOR THE PERIOD OF SAMPLING
NO
DATA IS VALID FOR ANY RECENT TYPICAL YEAR
SUMMARIZE DATA USING WIND2, STAR2 OR W NDROS
INPUT DATA INTO AIR QUALITY MODELS
PREPARE AIR QUALITY REPORT
END

STATISTICAL ANALYSIS FLOW CHART FOR AIR QUALITY SURVEY DATA


OBTAIN AIR QUALITY DATA FOR CORRELATION
WILCOXON OR REGRESSION TEST OF SAMPLING TECHNIQUE
SIGNIFICANT DIFFERENCE ?
YES
CHECK CALIBRATION OF BOTH INSTRUMENTS & RESOLVE PROBLEM
REOBTAIN SIDE-BY-SIDE AIR QUALITY DATA
NO
OBTAIN AIR QUALITY FIELD DATA
FRIEDMAN TEST FOR REDUNDANT AIR QUALITY SITES
SIGNIFICANT DIFFERENCE ?
NO
REDUCE NUMBER OF AIR QUALITY SITES
YES
OBTAIN REGRESSION EQUATION OF AIR QUALITY DATA
SIGNIFICANT RELATIONSHIP ?
YES
USE REGRESSION EQUATION TO ESTIMATE A.Q. CONC. WHEN NOT SAMPLING
NO
SUMMARIZE DATA USING HISTOGRAMS & SUMMARY TABLE
INPUT DATA INTO AIR QUALITY MODELS
PREPARE AIR QUALITY REPORT
END

NONPARAMETRIC STATISTICAL TESTS AND APPLICATIONS

NONPARAMETRIC STATISTICAL TESTS FOR AEROMETRIC DATA

2 SAMPLE CASE		K SAMPLE CASE	
RELATED COMPARING DATA FOR SAME TIME PERIOD	INDEPENDENT COMPARING DATA WITHOUT REGARD TO TIME	RELATED COMPARING DATA FOR SAME TIME PERIOD	INDEPENDENT COMPARING DATA WITHOUT REGARD TO TIME
SIGN TEST (MEASUREMENTS) WILCOXON MATCH SIGN TEST CORRELATION OF APCD VS DIV. OF HIGHWAYS. (MEASUREMENTS) '5; LSTAT; MPAIR'* RANDOMIZATION TEST (MEASUREMENTS)	MANN-WHITNEY U TEST COMPARING AIR OR METEOROLOGICAL DATA FOR RANDOM TIME PERIODS. (MEASUREMENTS) X^2 TEST (CONTINGENCY) STANDARD EXCEEDED (FREQUENCIES) '5; LSTAT; CONTIN'*	COCHRAN Q TEST (FREQUENCIES) FRIEDMAN 2-WAY ANOVA AIR QUALITY SURVEYS (MEASUREMENTS) '5; LSTAT; FRIED'* CATANOVA TEST WIND ROSE CORRELATION, STABILITY CORRELATION, CONCENTRATION EXCEED A STANDARD. (TIME PERIOD) (FREQUENCIES) '5; LSTAT; CATANOVA'*	X^2 TEST (CONTINGENCY) CORRELATE WIND ROSES FOR RANDOM TIME PERIODS, CONCENTRATIONS EXCEED A STANDARD FOR RANDOM TIME PERIOD. (FREQUENCIES) '5; LSTAT; CONTIN'* KRUSKAL-WALLIS TEST COMPARING AIR QUALITY OR METEOROLOGICAL DATA FOR A RANDOM TIME PERIOD. (MEASUREMENTS)

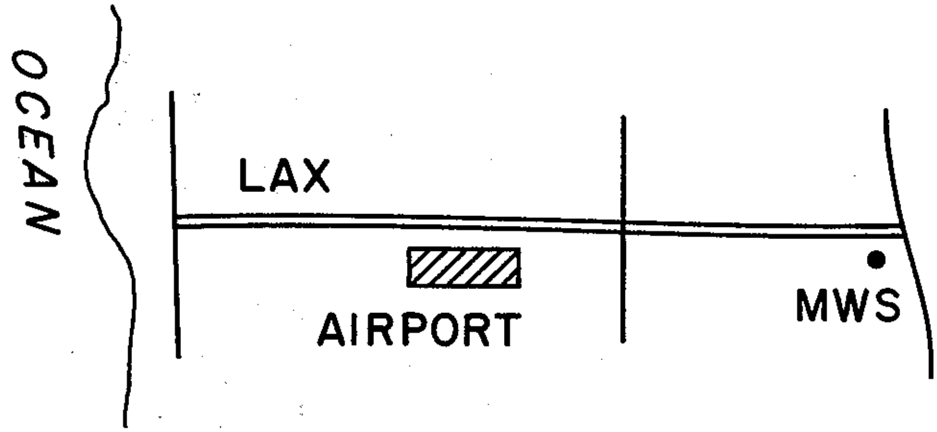
MEASUREMENTS = ACTUAL DATA
FREQUENCY = NO. OF OCCURANCES

CONFIDENCE LIMITS – USE FOR SPECIFIED TIME PERIOD CONSISTENT WITH MATHEMATICAL MODEL PREDICTIONS.
'5; LSTAT; CONLIM'*

TOLERANCE LIMITS – GENERAL SUMMARY OF ALL DATA COLLECTED (AIR QUALITY) FOR ENTIRE SAMPLING PERIOD (INDEPENDENT OF TIME).

*PROGRAMS AVAILABLE ON TENET TIME-SHARE COMPUTER

EXAMPLE NO. 14: Wind Rose Correlation (Catanova Test)



Purpose:

1. Reduce the use of mechanical weather station (MWS).

2. Use one station for the meteorological inputs into mathematical model.

3. In general determine redundant information statistically.

Compare wind roses for all stations for July 0700 - 0900

1. Wind speed

2. Wind directions

H_o: There is no significant difference in the distributions of wind speeds and directions for all three meteorological sources between the hours of 0700 - 0900 for the month of July.

H_A: H_o is not true.

DIRECTION	LAX	AIRPORT	MWS
N			
NNE			
NE			
ENE			
E			
ESE			
SE			
SSE			
S			
SSW			
SW			
WSW			
W			
WNW			
NW			
NNW			
Calm			

Obtain the occurrences for directions from computer program outputs. (See Pg.118)

H_{o1}: $(\emptyset)_{LAX}=(\emptyset)_{AIR}(\emptyset)_{MIR}$

H_{A1}: $(\emptyset)_{LAX}=(\emptyset)_{AIR}=(\emptyset)MRI$

Catanova test requires frequency distributions of wind directions.

WIND SPEED	LAX	AIRPORT	MWS
0-3 mph			
4-7 mph			
8-12 mph			
13-18 mph			
19-24 mph			
> 24 mph			

Obtain wind speed occurrences from computer program outputs. (See Pg. 118)

H_{o2}: $(\overline{U})_{LAX}=(\overline{U})_{AIR}=(\overline{U})_{MRI}$

H_{A2}: $(\overline{U})_{LAX}=(\overline{U})_{AIR}=(\overline{U})_{MRI}$

Both H_{o1} and H_{o2} must be true to accept H_o.

EXAMPLE NO. 15: Comparison of Stability Classes (Catanova)

Same as Example No. 14 (two metheorological sources). Compare stability classes for December 1972, January 1973 and February 1973 for 0700-0900 hrs. Assume cloud cover and ceiling height measured at both airports.

Purpose: Determine if monthly data can be combined in computer program for the winter season.

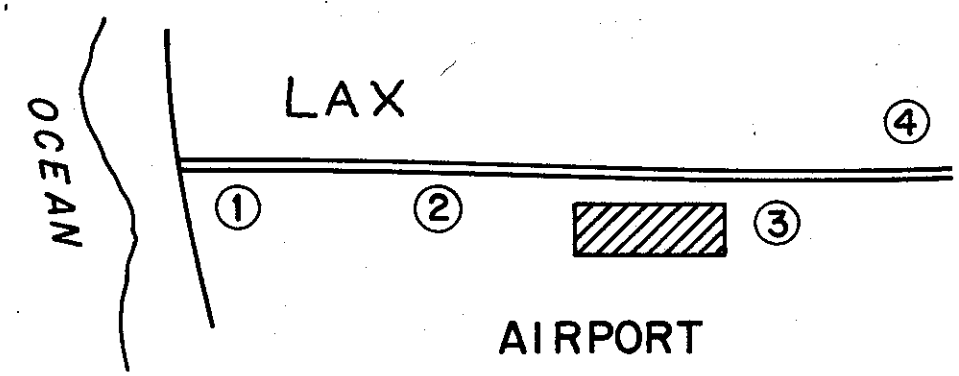
H_o: There is no significant difference between the distribution of the Stability Classes A through F for December, January and February for 0700-0900 hrs. for LAX and Airport.

H_A: H_o is not true.

	LAX	AIRPORT
A		
B		
C		
D		
E		
F		

Obtain occurrences from computer programs WNDROS or STAR2.

EXAMPLE NO. 16: Air Quality Survey (Friedman 2-Way ANOVA)
Spatial Distribution



Purposes:

1. In general statistically determine the redundant sampling sites.

2. Eliminate redundant information as inputs into mathematical model.

3. Reduce manpower requirements.

H_o: There is no significant difference in the spatial distribution of CO for all sites between the hours of 0600 to 1000.

H_A: H_o is not true.

Spatial Distribution

TIME	SITE 1	SITE 2	SITE 3	SITE 4
0600-0700				
0800-0900				
0900-1000				

Rank data among sites.

Must also apply Friedman 2-way ANOVA for mid-day and evening hours or other time periods of interest before eleminating a site.

Why?

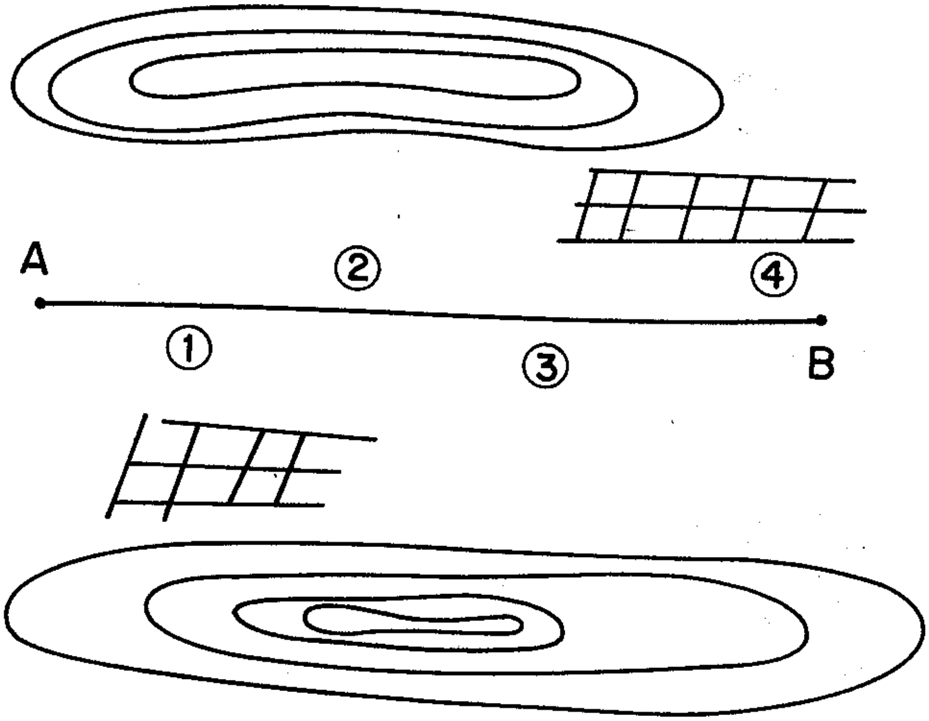
For rural highway projects this is a valid approach to justify sampling at fewer sites and thus reduce costs provided the analysis was made on the entire daily sampling. The same applies for urban projects as well. However, if area is politically or environmentally sensitive you may want to consider monitoring in all hours regardless of statistical outcome and costs.

For a rural project where the maximum values are generally small (≤ 4 ppm), analyze the data using Friedman test for the entire day. If no significant difference is obtained there is justification for not sampling every site. This can minimize field costs. Special consideration should be given to monitor the sites that measure high values.

EXAMPLE NO. 17: Air Quality Survey Friedman 2-Way ANOVA Temporal Distribution

Given:

A "Blanket" two week quality survey was made from 0600 to 1900 hours.



Determine:

1. If it is necessary to sample every site.

2. If it is necessary to sample every consecutive hour.

3. Re-evaluate and design an adequate air survey to meet objectives above if possible.

Solution:

A. First must determine that there is no significant difference in CO spatial distributions at all sites for the daily sampling period (Example No. 16). If there is a significant difference must monitor all sites based on the analyses. If not, consider the temporal distribution as discussed below.

B. Next, apply Friedman 2-way ANOVA as follows:

H_o: CO concentrations do not change significantly with time for Sites 1 through 4 for the day(s) sampled.

H_A: H_o is not true.

Temporal

Time (Days)

Site	0600-0700	0700-0800	0800-0900	--- ---	1700-1800
1					
2					
3					
4					

Rank data among sites

If you accept the null hypotheses (H_o) in Parts A and B above using the Friedman 2-way ANOVA, then there is justification for not sampling at every site for every hour of the daily sampling period.

Example:

Sample 2 hour during AM and 2 hours PM etc. The revised sampling plan depends on manpower, equipment available, and sensitivity of project.

This is a good approach for rural projects to reduce sampling. However, one must consider seasonal variations in meteorology that influence air quality. This analysis should be made for different meteorological conditions to assure that the temporal and spatial distributions do not change.

EXAMPLE NO. 18: Comparison of APCD Continuous Monitoring vs. Bag Sampling (Wilcoxin Matched Sign Test)

H_o: There is no significant difference in APCD one-hour average concentration vs. Bag Sampling for one-hour.

$$(CO)_{APCD} = (CO)_{BS}$$

H_A: H_o is not true.

$$(CO)_{APCD} \quad (CO)_{BS}$$

TIME	APCD	BAG SAMPLING
0700-0800		
0800-0900		
0900-1000		
1000-1100		
1100-1200		
1200-1300		
1300-1400		
1400-1500		

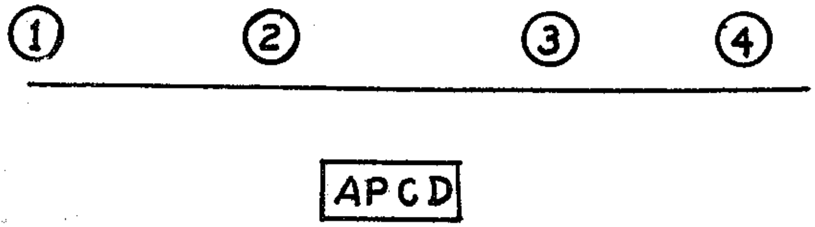
If the difference in CO comparison above is ± 1 ppm, for the "real life world" there is no significant difference regardless of the statistical outcome.

For this type of analysis it is very important to assure that both APCD and Caltrans calibrate their equipment before the comparison is made to avoid improper interpretation of results.

It is recommended that this comparison test be made at the beginning of the sampling period, midway, and at the end of monitoring for a complete consistency check. This test should cover peak AM and PM traffic hours along with off peak traffic to compare high and low concentrations.

REGRESSION ANALYSES

EXAMPLE NO. 19: Prediction of Pollutant Concentrations



Purpose:

Random sampling every 3-days. Regression techniques will allow prediction of pollutant concentrations at sites for days not sampled.

$(CO)_1$ = a+b. (APCD) (Simple linear regression)

$(CO)_1 = a+b.\ (APCD) + c\ (\frac{1}{\bar{U}})$

$(CO)_1 = a+b.\ (APCD) + c\ (\frac{1}{\bar{U}}) + d\ (\text{ceil. ht.}) + e\ (\text{cloud cover})$

try log transforms. Why?

a, b, c, d, e = regression coefficients.

Computer programs will determine regression coefficients.

It may be desirable under certain circumstances to include time, "t" as an independent variable.

$$(CO)_1 = a + b\ (APCD) + C\ (\frac{1}{\overline{U}}) = d(t)$$

It must be emphasized that the boundary condition for the data base used to determine the regression coefficients must not be exceeded in future predictions.

EXAMPLE NO. 20: Wind Correlation

MWS
①
② AIRPORT

Procedure:

1. Use CATANOVA test to see if there is a statistical difference in ø and $\overline{U}$. If no difference, there is no need to use the regression techniques.

   Accept H_O

2. Accept H_A. Use regression techniques.

Wind Speeds:

$\overline{U} = a + b\overline{U}_2$

$\overline{U} = a + b\overline{U}_2 + cH$ H = inversion base

$\overline{U} = a + b\overline{U}_2 + cH + d$ (cloud cover) + e (ceiling height)

Wind Directions

$ø_1 = a + bø_2$

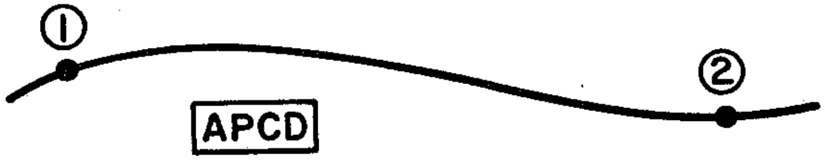
$ø_1 = a + bø_2 + cH$ H = inversion base

$ø_1 = a + bø_2 + cH + d$ (cloud cover) + e (ceiling height)

try log transforms

This regression equation can possibly explain the difference in pollutant concentrations when $\overline{U}$ and ø is used in the mathematical model. Also to explain why background pollutant levels change from location 1 to location 2.

EXAMPLE NO. 21: Air Quality Survey - Mini-Van Random Sampling



Purpose: Develop prediction equation for O_3 for days not sampled.

This approach may be especially useful if a high O_3 concentration is measured at the APCD station when not sampling at the site.

$$O_3 = a + b\ (APCD)$$

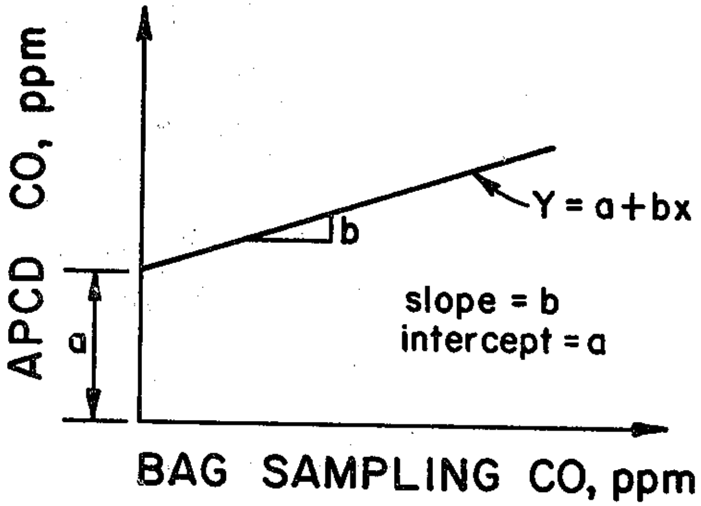
$$O_3 = a + b\ (APCD) + c\ \left(\frac{1}{\overline{U}}\right)$$

$$O_3 = a + b\ (APCD) + c\ \left(\frac{1}{\overline{U}}\right)\ dH$$

Try log transforms. Why?

Time may also provide a better prediction for O_3 because of the diurnal variation of O_3.

EXAMPLE NO. 22: Correlate APCD Sampling and Bag Sampling



For a perfect correlation (1 to 1) the line of best fit is a 45 degree line passing through the origin.

Procedure:

1. $t = \frac{b-b_o}{S_b}$ H_o: b = 1 (45° line)

2. $t = \frac{a-a_o}{S_a}$ H_o: a = o (Zero intercept)

If a ≠ o then a systematic error.

Note: 1. If the range of data is less than ± 4 ppm or the number of observations is less than 15, regression analysis may give misleading results. Try non-parametric test (Wilcoxin Matched Sign Test). See Example No. 18.

CHI-SQUARE TEST

GOODNESS-OF-FIT TEST

In many practical applications, we often assume that the population data follows a theoretical distribution such as normal or lognormal. Any assumed theoretical distribution for the population may be questioned. A method for checking the validity of this assumed distribution is the Chi-square goodness-of-fit test.

This test is used to determine how well an assumed theoretical distribution fits an empirical distribution obtained from sample data. The procedure is to make a comparison between the observed frequencies from the sample and the expected frequencies under the "assumption" and to measure how much discrepancy exists between them.

By following the general test procedure, outlined in the "Hypothesis Testing Section" we have the following:

1. State the experimental goal.

2. State the statistical null hypothesis.

3. Choose the level of significance, α, the chance of rejecting the null hypothesis when it is actually true.

4. Choose an appropriate statistical test for testing the null hypothesis.

    (the test statistic for the Chi-square goodness-of-fit test is Chi-square)

$$\chi^2 = \sum_{i=1}^{b} \frac{(O_i - e_i)^2}{e_i}$$

Where O_i = the observed frequency in the i^{th} class.

e_i = the expected frequency in the i^{th} class, calculated according to the assumed theoretical distribution.

$i = 1, 2, \ldots, b$ classes or events

5. The sampling distribution is the Chi-square, with df = b-1-k, where k is the number of parameters used in estimating the expected frequency.

6. The rejection region is $\chi^2 \geq \chi^2_{1-\alpha, df}$

7. Compute the test statistic value from a random sample.

8. If computed χ^2 falls into the rejection region, reject H_o; if not, accept H_o.

9. State the experimental conclusion.

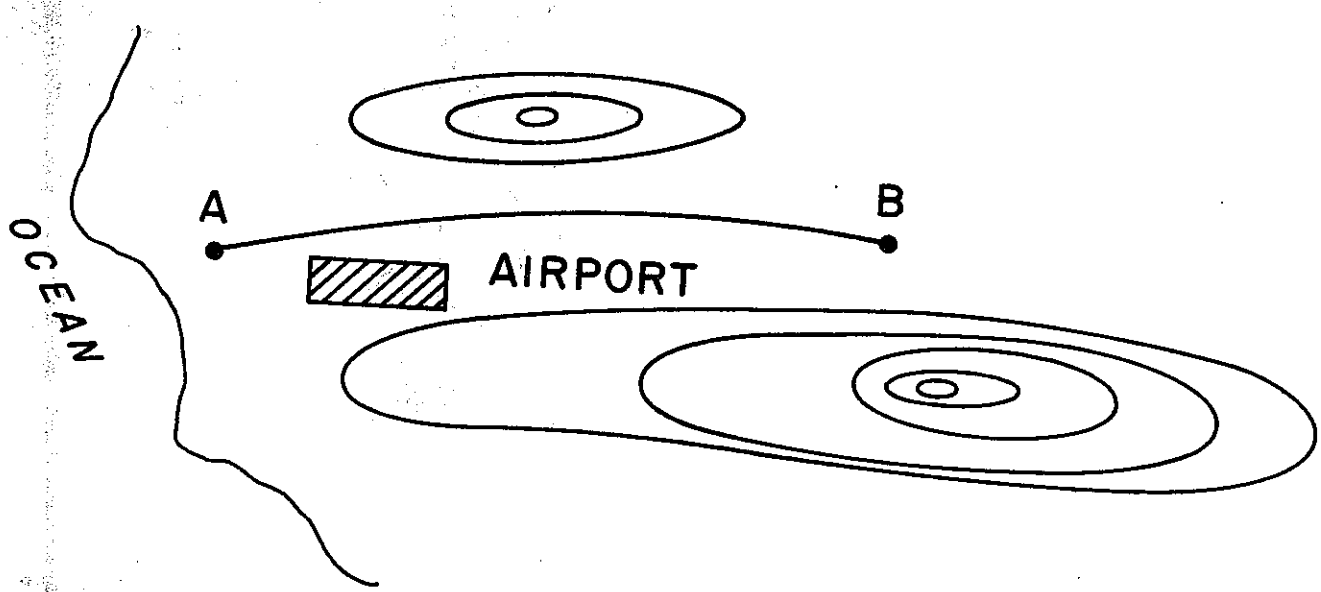
DETERMINATION OF TYPICAL YEAR

Purpose:

1. To determine if a sampling period (year, month, etc.) is representative of the past historical records (years, months, etc.).

2. Define a typical year in terms of meteorology or smog season.

Example 23:



Given:

Ten years of historical meteorological data is available at the airport on a continuous 24 hour basis. For the proposed route AB it was determined that an additional weather station was needed to supplement the existing data at the airport to determine the temporal and spatial distributions of the surface streamlines. The weather station was in operation for a one year period. Local newspaper reports and environmentalist indicate that the year the MWS was in operation was not a representative year because wind speeds were generally stronger than past years. They also indicate that the weather front movements in the region changed significantly over the past year, changing the general surface wind direction.

H_o: Determine if the frequency distribution of wind speeds and directions measured at the airport are representative of the past ten years of historical records.

H_A: H_o is not true.

Solution:

Use of Chi-square test.

Since air quality changes seasonally (summer O_3; winter CO, HC, NO_x) it is best to analyze $\overline{U}$ and ϕ on a seasonal or monthly basis. For this analysis use data at airport.

1. For O_3 season (summer and fall) O_3 most critical 1200 - 1600.

Compare:

$$\left.\begin{array}{l}(\overline{U})\text{ summer vs }(\overline{U})\text{ historical summer}\\[1em](\varnothing)\text{ summer vs }(\varnothing)\text{ historical summer}\end{array}\right\}\begin{array}{l}\text{Seasonal or}\\\text{monthly for}\\\text{specified hours}\end{array}$$

2. For primary pollutants (CO, HC, NO_x) most critical AM and PM during winter season.

Compare:

$$\left.\begin{array}{l}(\overline{U})\text{ winter vs }(\overline{U})\text{ historical winter}\\[1em](\varnothing)\text{ winter vs }(\varnothing)\text{ historical winter}\end{array}\right\}\begin{array}{l}\text{Specify hours}\\\text{and season or}\\\text{months}\end{array}$$

In order to be absolutely certain that a "typical" year was observed, all of the following variables must be examined:

1. Wind speed
2. Wind direction
3. Stability
4. Inversion base height - soundings or aircraft measurements.
5. Incoming solar radiation (the temperature can be used as an estimate)

If the frequency distribution of all the variables are not significantly different from their historical counterparts, then a typical year was observed.

It may not be possible to obtain all the above data. Consider as many of the above variables as possible. Be sure to indicate what variables were used. This discussion should be included in the portion of the report dealing with meteorology.

Obtain wind speed frequency, and wind direction frequency, data from the computer printout of programs WIND2, STAR2, or WNDROS. Consider the overall wind rose (class 9) because we are concerned with the general meteorological condition over a large time period. Stability frequency data is obtainable only from STAR2 or WNDROS.

Data on incoming radiation may be available from local APCD's. If not, the temperature can be used to estimate this parameter. However this is not a good estimate if either the aerosol content or the water content of the atmosphere is high.

DETERMINATION OF WORST AND TYPICAL BACKGROUND LEVELS

INTRODUCTION

The determination of the highest expected yearly pollutant concentrations for any averaging time will typically involve the use of Larsen's model. Larsen's model extrapolates the maximum yearly concentration for a pollutant-concentration distribution with a one year base to that of a known distribution with a smaller base (for a constant averaging time). Secondly, the model allows conversion from a known distribution for a particular averaging time to a distribution for any other averaging time for comparisons with air quality standards.

Specifically, Larsen's model can be used for the following purposes:

1. For carbon monoxide especially, using observed field data:

   a. To predict the maximum annual expected one-hour concentration for the present year, the estimated time of completion (ETC), and the estimated year of completion + 20 years (ETC+20).

   b. Predict 8 and 12 hour annual maximum expected carbon monoxide concentrations both in the microscale and mesoscale region under various alternates for transportation schemes (i.e., various projects and "no-build" alternatives).

2. Make maximum utilization of observed field data in prediction methods for ambient background concentration values.

3. For photochemical pollutants such as NO_2, O_3, and HC where a photochemical simulation model is not used or available, and sampled data is available, to predict ambient background values based upon the sampled data and to predict ambient concentrations under different alternatives.

4. Utilization of "scarce" data and APCD historical data to predict ambient concentrations.

5. To predict frequencies of occurrence of maximum concentrations for any averagint time. The number of exceedances of any standard can also be predicted for any averaging time for any pollutant.

Reasons for using Larsen's Model.

1. Larsen's Model is an EPA recognized mathematical model of air quality measurements and is an EPA publication. Dr. Larsen works with the EPA at Research Triangle Park, North Carolina, and is the author of AP-89, A Mathematical Model Relating Air Quality Measurements to Air Quality Standards.

2. Larsen's model provides an easily handled mathematical model for predicting pollutant concentrations. It is based upon the assumption that all air quality pollutant concentration measurements are log-normally distributed. Although some sets of data depart from lognormality, use of the lognormal distribution is recommended because most aerometric pollutant data tend to fit this distribution better than any other.

3. Although Larsen's model may not prove to be the ultimate tool for predicting ambient concentrations of aerometric pollutants, it is now acceptable, available, useable, and verified,

especially for urban areas, and appears good also for rural areas. (There is no definition of an urban area vs. a rural area, but preliminary indications in California indicate that it holds practically everywhere.)

4. To demonstrate future applications and utilization of Larsen's model for sampling procedures, precision of sampling, confidence bands of data sampled and predicted, and the reliability and accuracy of the measured and presented data in an air quality impact report.

Larsen's model is the result of years of work by Larsen and his associates at EPA. The earlier versions of the model as discussed in these notes can be followed in the technical journals for over the last ten years. The model was verified and compared with measured values for seven gaseous pollutant concentrations obtained during continuous sampling for up to seven years in eight cities to obtain methods for predicting concentrations for various averaging times. Two of these cities were Los Angeles and San Francisco. The student should become familiar with the text of Larsen's Report (AP-89).

Larsen Model Characteristics:

The assumptions inherent in Larsen's Model are the following:

1. Pollutant concentrations are log-normally distributed for all averaging times (i.e., cumulative percentile data plot as a straight line on log-normal probability paper).

2. Median concentrations are proportional to averaging time raised to an exponent (i.e., they can be plotted as a straight line on log paper).

3. The arithmetic mean concentration is the same for all averaging times.

4. Maximum concentrations are approximately inversely proportional to averaging time raised to an exponent.

5. For the longest averaging time calculated (usually 1 year), the arithmetic mean, geometric mean, maximum concentration, and minimum concentrations are all equal.

Using the above assumptions, Larsen developed equations to predict pollutant parameter characteristics regardless of the averaging time used.

The required statistical parameters are:

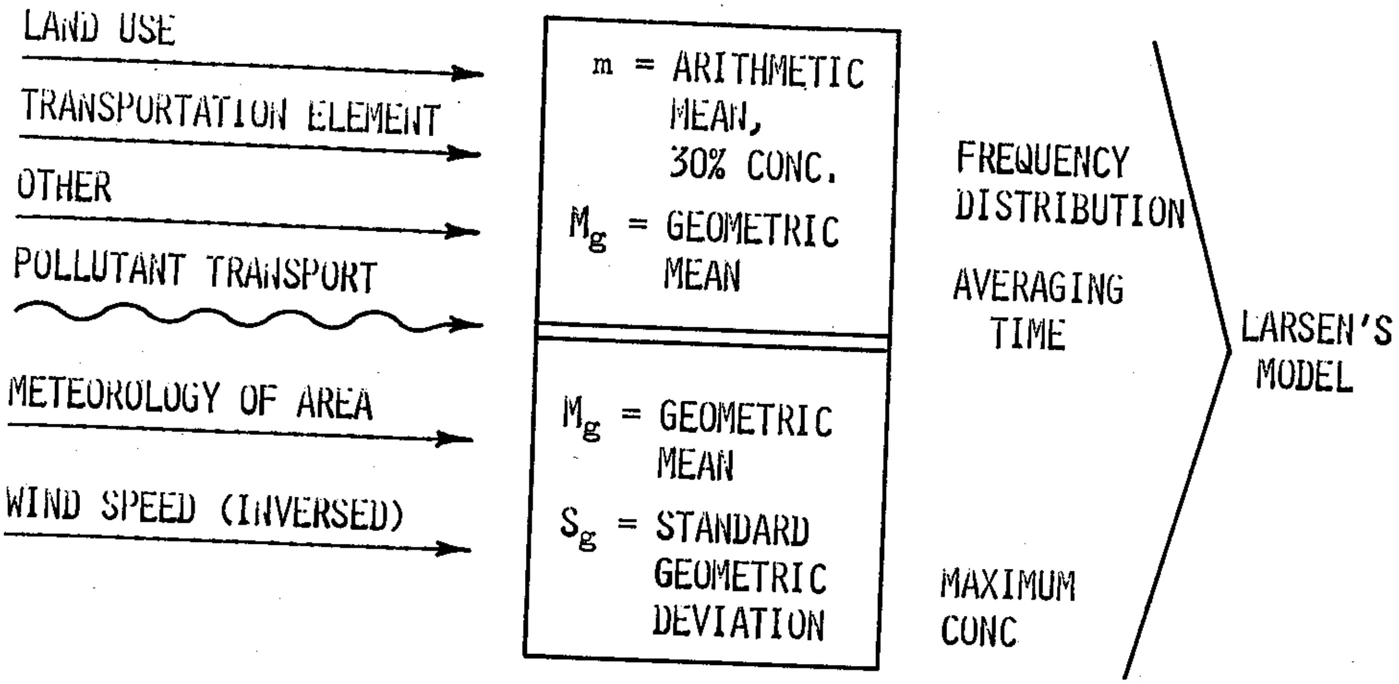
1. Geometric mean, m_g, or arithmetic mean, m

2. Standard geometric deviation, s_g

3. Maximum concentration expected once-a-year for a particular averaging time, c_{max}

4. Frequency distribution of expected pollutant concentrations, (i.e., log-normal).

When utilizing Larsen's Model, the mean value, m, can be thought of as the representation of the pollutant burden or some proportion thereof, which is relatively constant over a period of time (perhaps a year or two). The standard geometric deviation, s_g, can be thought of as a function of the meteorology of the area, most likely closely associated with the reciprocal of the wind speed as a related parameter. Perhaps the diagram shown below can help explain Larsen's Model and the various elements of the model which we would normally use.



Although Larsen derived his model by empirical methods and enormous data analysis, other researchers, notably Pollack, Kahn, and Gifford have separately shown mathematical derivations of the log-normal distribution for air pollutant concentrations. Larsen's model is unique in its averaging time analysis. Other averaging time analysis techniques have been suggested by Turner and Hino, but will not be discussed in these notes.

Examples

On the following pages are examples of how Larsen's Model can be used in the assessment of air quality impact of transportation projects.

Data Presentation Using Larsen's Model and AP-89

It is suggested, as a minimum, that when using Larsen's Model the following be included in the data presentation:

1. Tabular summary to include:

   a. distribution of observed data--ranks, categories, etc.

   b. arithmetic mean, (m)

   c. geometric mean (m_g), with confidence interval if available

   d. number of samples and sample dates and times of samples

   e. standard geometric deviation (s_g)

2. Graphical summary, to include:

   a. histogram or frequency curve

   b. plot of data on log-probability paper using cumulative frequency

      1) show m_g, s_g, and maximum concentration for averaging times of interest

3. If hourly data is presented for CO for instance, predict 8 and 12 hour averaging times concentrations and give comparisons to observed data of same time period(s) where available.

For example the following data were observed in the field in the winter of 1972-1973 during December, January and February: 55 hours @ 1 ppm; 148 hours @ 2 ppm; 65 hours @ 3 ppm; 45 hours @ 4 ppm; 15 hours @ 5 ppm; 3 hours @ 6 ppm. (This is actual field data.)

ppm	No./Occ.	Rank	Cum%	Gumbel%	f %
6	3	1-3	.91	.90	.181-.79
5	15	4-18	5.43	5.42	1.09-5.32
4	45	19-63	19.03	18.98	5.62-18.91
3	65	64-128	38.67	38.55	19.34-38.55
2	148	129-276	83.38	83.13	38.97-83.38
1	55	277-331	100.0	99.70	83.69-99.82

$\Sigma = 331$

Example 23a:

Hand calculations using Larsen's Model:

Using Larsen's methodology for non-continuous data (Reference Page 32, AP-89).

Arithmetic Mean = 2.48

Max Observed Value = 6 PPM
Ranks 1-3
Median Rank = 2

$$f = \frac{100\ (r-0.4)}{n} = \frac{100(1.6)}{331} = 0.4833$$

$A(z) = 0.4952 \quad z = 2.59$

$$
\begin{aligned}
S_{gi} &= \exp\left\{z-[Z^2-2\ \ln \tfrac{C}{M}]^{0.5}\right\} \\
&= \exp\left\{2.59-[2.59^2-2\ \ln \tfrac{6}{2.48}]^{0.5}\right\} \\
&= \exp\left\{2.59-[6.72-2(0.884)]^{0.5}\right\} \\
&= \exp\left\{2.59-2.22\right\} \\
&= \exp\left\{0.37\right\} \\
S_{gi} &= 1.45 \\
M_g &= C/S_g^z = {}^{6}/1.45^{2.59} = {}^{6}/2.62 = 2.29 \text{ PPM}
\end{aligned}
$$

(from $C = M_g S_g^z$

Confidence bands about median or geometric mean from prior notes - (Saltzman Chart)

95% Confidence bands ±4% = ± .09 PPM;

99% Confidence bands ±5% = ± .11 PPM

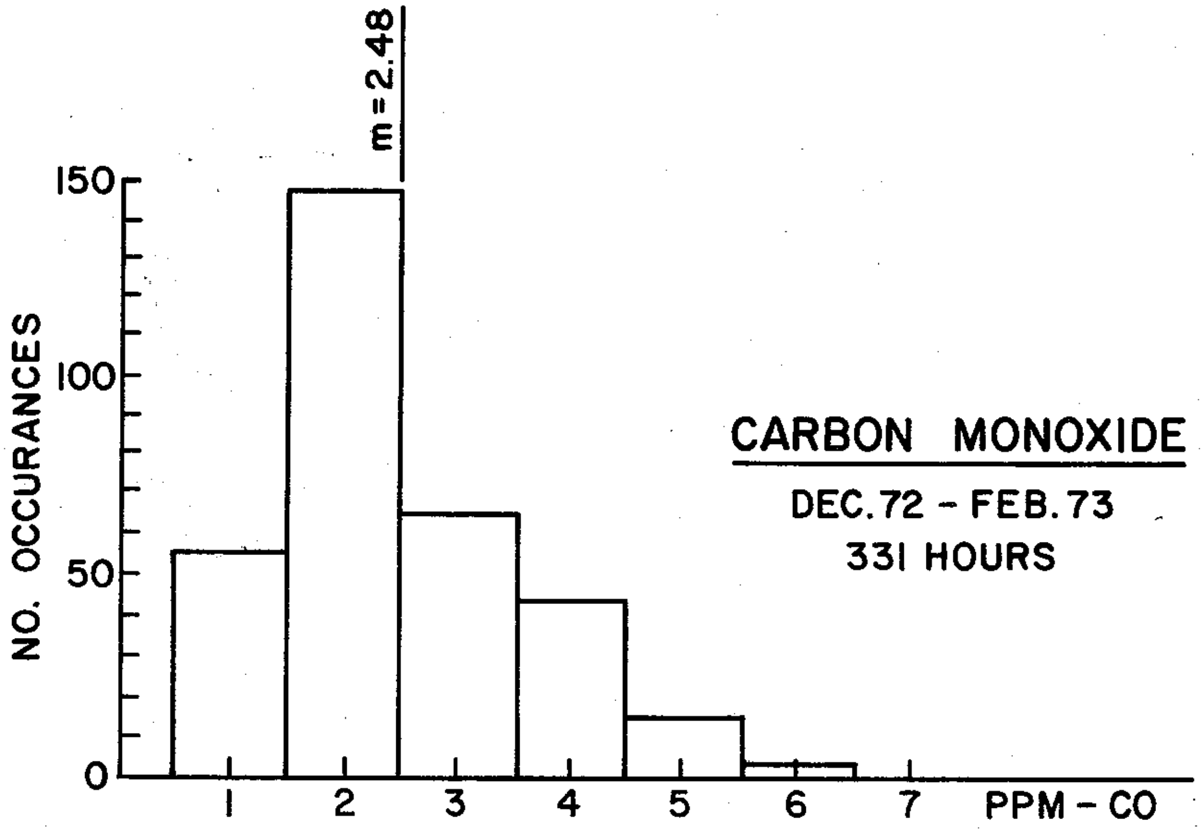
$C_{max\ hr} = 2.29\ \underbrace{(1.45)^{3.81}}_{4.12} = 9.45$ PPM ⇐ Report only significant digits but hold one extra place during computations

$S_{g8} = 1.39$ (Table 14, AP-89)

$C_{max8} = (2.75)\ (\overset{M}{2.48}) = 6.82$ PPM (Table 14, AP-89)

$M_{g8} = 6.82/\underbrace{(1.39)^{3.26}}_{2.92} = 2.34$ PPM

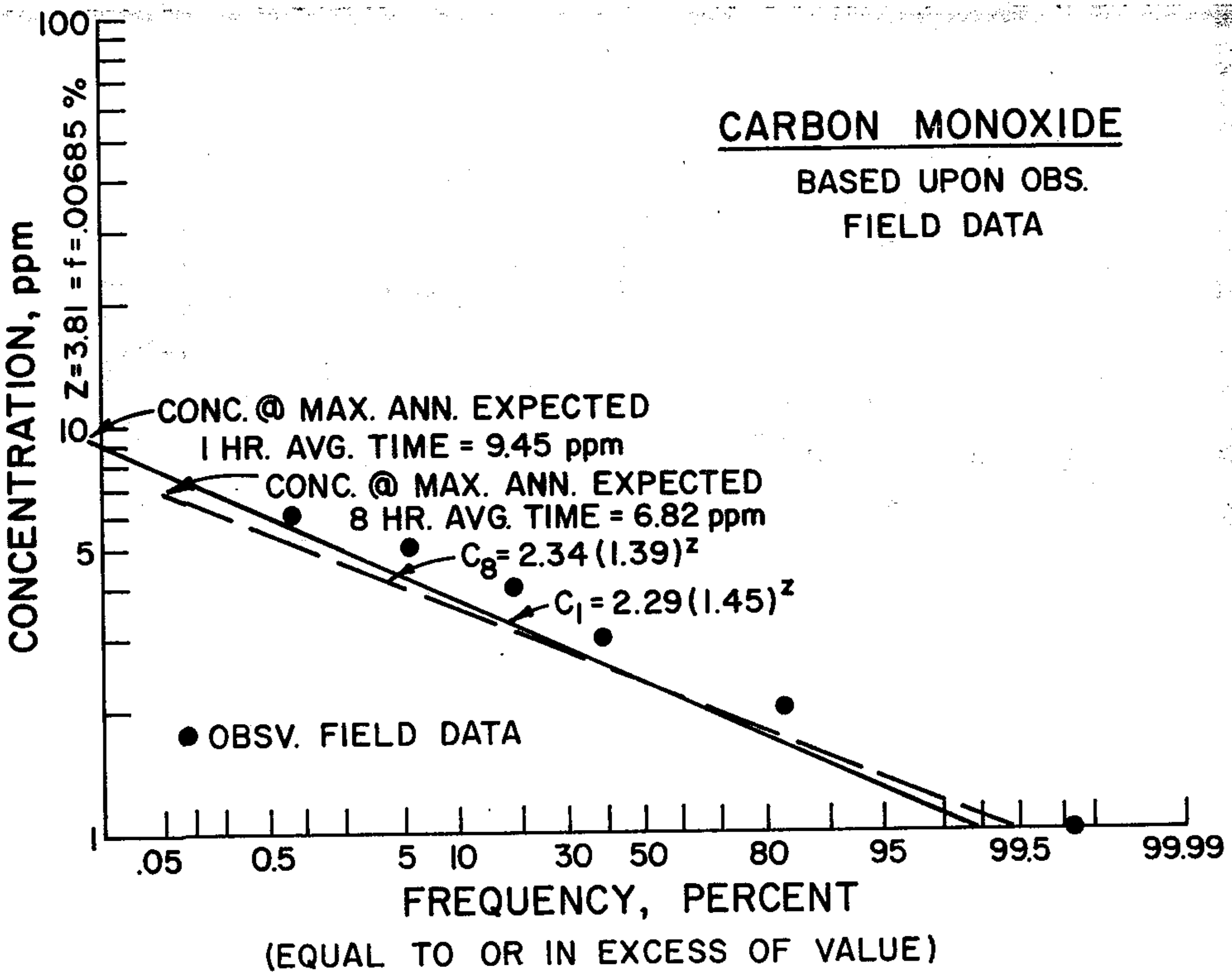
Z from table #11, AP-89



CARBON MONOXIDE

DEC. 72 - FEB. 73

331 HOURS


CARBON MONOXIDE
BASED UPON OBS.
FIELD DATA
CONCENTRATION, ppm
Z=3.81 = f=.00685 %
100
10
5
1
CONC. @ MAX. ANN. EXPECTED
1 HR. AVG. TIME = 9.45 ppm
CONC. @ MAX. ANN. EXPECTED
8 HR. AVG. TIME = 6.82 ppm
$C_8 = 2.34(1.39)^z$
$C_1 = 2.29(1.45)^z$
● OBSV. FIELD DATA
.05
0.5
5
10
30
50
80
95
99.5
99.99
FREQUENCY, PERCENT
(EQUAL TO OR IN EXCESS OF VALUE)

Example 24:

Comparing field site to APCD -- Short Duration Sample

Presented here is a method for estimating yearly frequency distributions of pollutants at a project site or corridor based upon limited sampling on the site, but where an APCD station is nearby and a comparison can be made of the observed concentrations on the project site and the APCD station at the same exact times. The methodology suggested appears useable for several pollutants including oxidants and carbon monoxide. It is suggested for use where no better information is available and time does not permit extensive sampling.

The basic advantage to this method is the possibility of shortening the sampling time by pertinent observations. The following steps are suggested to implement the method:

1. Sample at the site representative of the project for a month or longer, preferably longer. Sampling need not be continuous but should be done with some high degree of frequency. A high percentage of the available hours in any time period should be sampled to increase reliability.

2. Compare the standard geometric deviation and the arithmetic mean of the project site with the nearest APCD for exact same period. Use Larsen's Model with non-continuous sampling to solve for s_g, m_g, and m. Always plot the results of the sampling as simple superposition may show the relationship of the site to the APCD station and further calculation may be unnecessary.

3. Compare the site and APCD s_g and m and calculate ratios between the site and the APCD for the observed data sampling time. Using these ratios and past data of the APCD station

s_g, m_g, and m calculate the estimated annual frequency distribution of the pollutant at the site. The assumption used in this method is that the s_g difference should not be too great, i. e., similar meteorology exists and seasonal meteorological variations are small between site and APCD.

The following hourly CO data were obtained at a project location during the month of December:

One hour CO samples obtained in December @ Site #1 and APCD station.

	Site #1		APCD	
	Year	Month	Year	Month
m	---	3.5	*3.0	4.0
s_g	___	1.7	*1.6	1.5

*Denotes available historical data.

1. Plot the data (use MATHISTO computer program) to test for lognormality.

2. Site #1's annual arithmetic mean, m, and standard geometric deviation s_g, can be estimated as follows:

$$m = (3.5)\ (3.0)/4.0 = 2.62 \text{ ppm}$$

$$s_g = (1.7)\ (1.6)/1.5 = 1.81 \text{ ppm}$$

This is just straight proportioning.

The maximum expected 1-hour CO concentration can be estimated as follows:

Using equation #11 on Page 10 of AP-89:

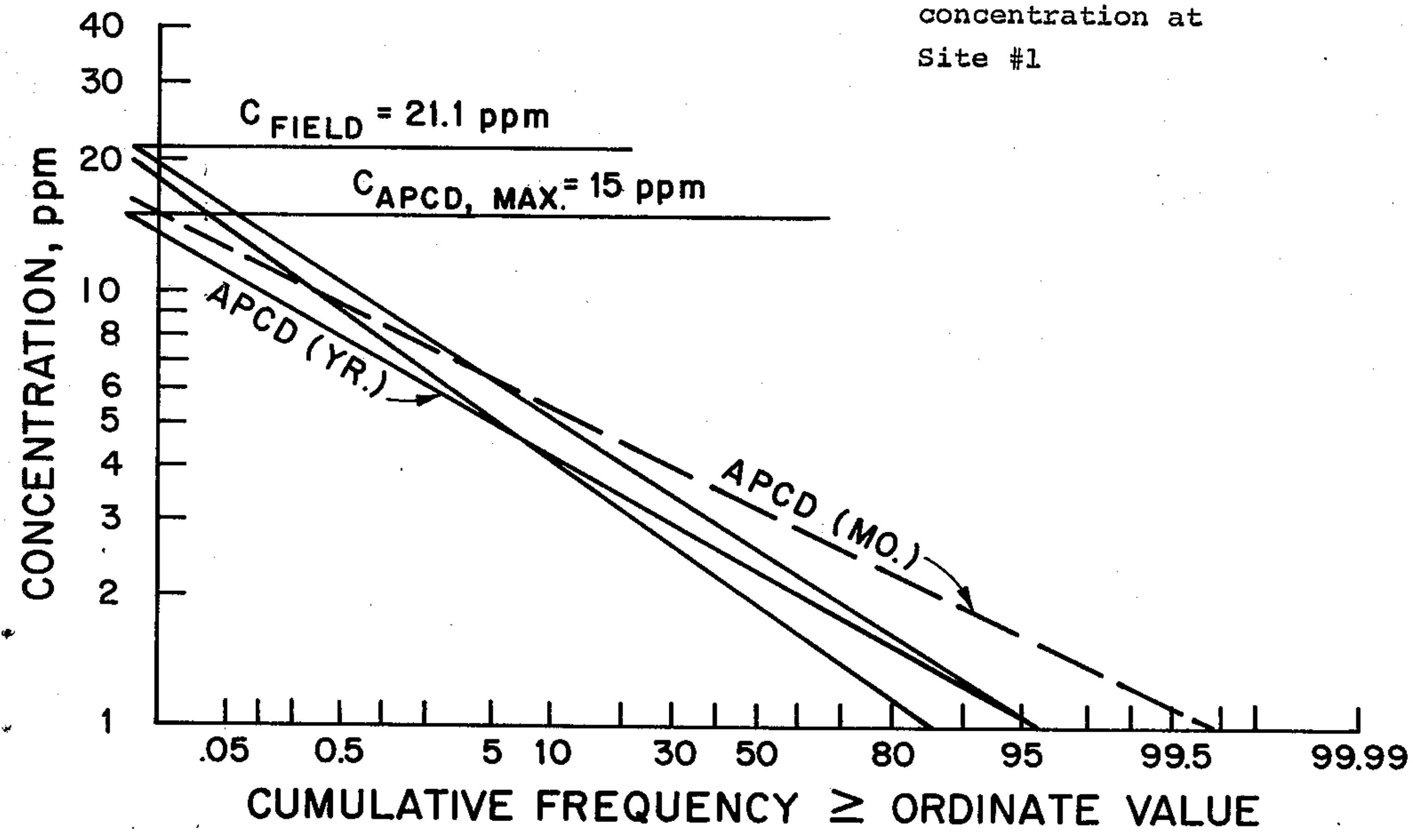
$$m_g = m/\exp(0.5\ \ln^2 s_g)$$

$$= 2.62/\exp(0.5\ \ln^2 1.81)$$

$$m_g = 2.20$$

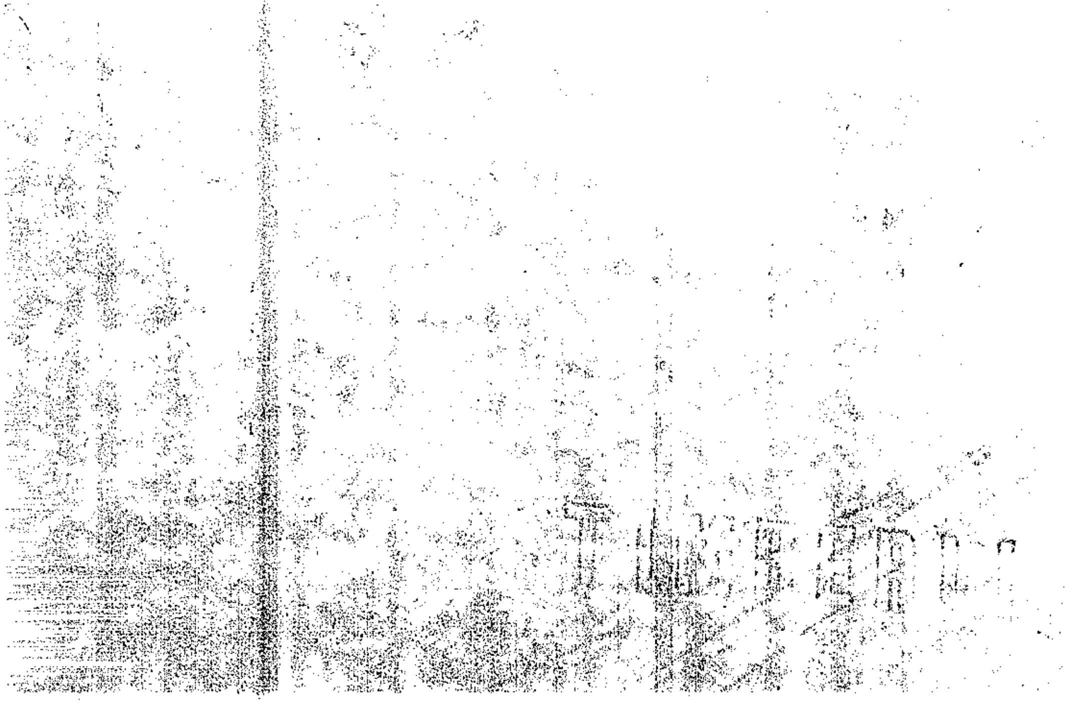
$$c = m_g s_g{}^z$$

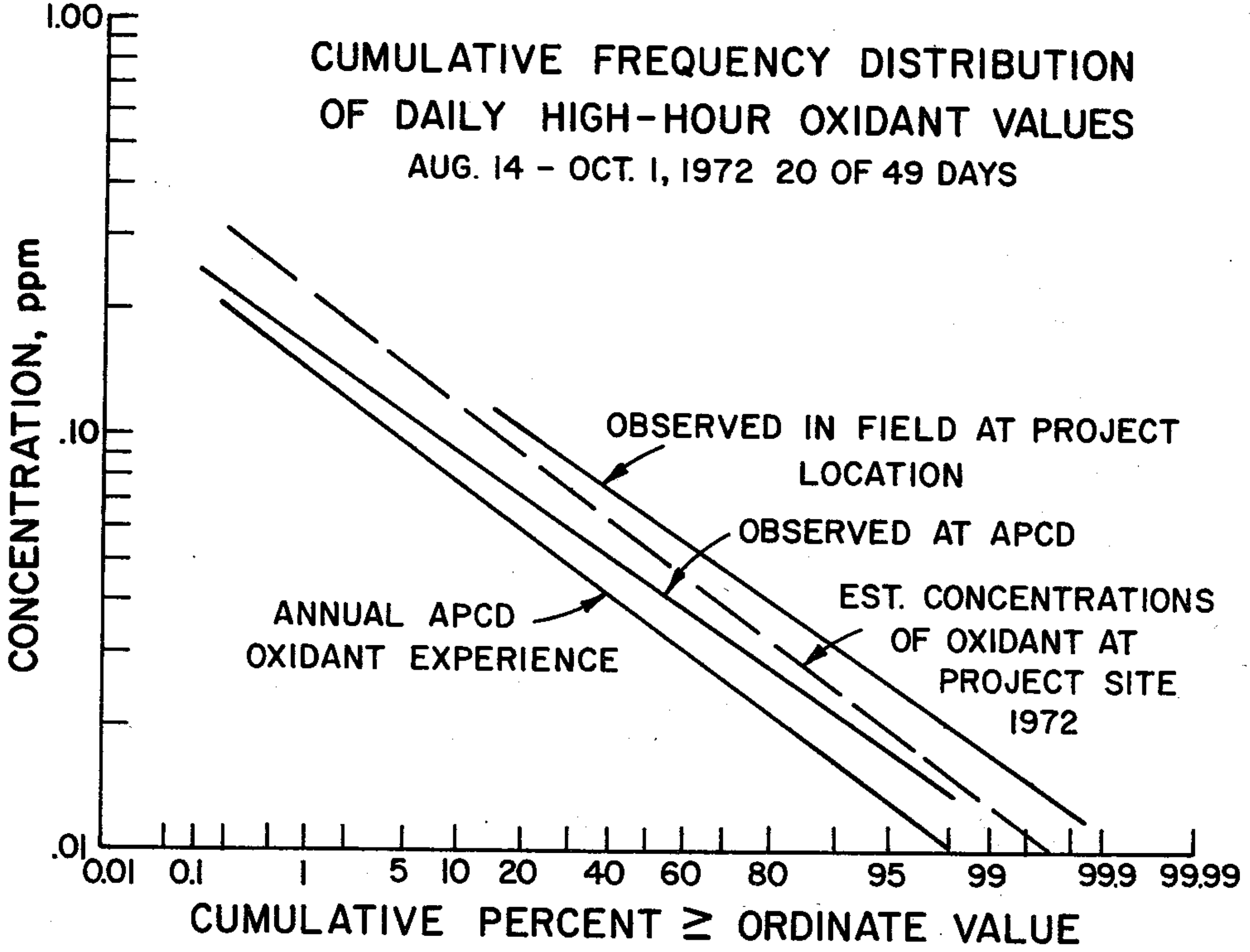
$c_{max} = (2.20)\ (1.81)^{3.81} = 21.1$ ppm---annual expected one-hour maximum concentration at Site #1



This methodology takes advantage of the Larsen's Model assumption that the averaging time arithmetic mean is the same for all averaging times and that the one-hour arithmetic mean approximates the 30 percentile for a lognormal distribution. The arithmetic mean is a function of land use and this should be accounted for in any future calculations of emission concentrations (tons/day analysis).

A graphical solution to a similar problem is shown below. Here the same type of data is shown graphically and the entire results were done graphically, especially since the field data and APCD data have such similar s_g's.



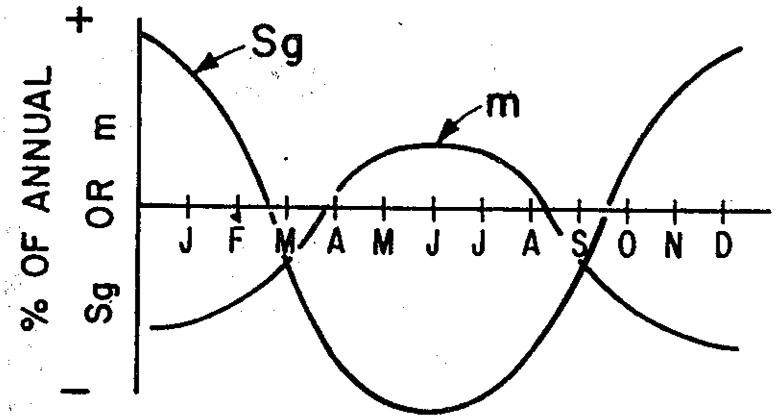


CUMULATIVE FREQUENCY DISTRIBUTION OF DAILY HIGH-HOUR OXIDANT VALUES

AUG. 14 - OCT. 1, 1972 20 OF 49 DAYS

This method can be used to determine when to sample, and to relate the relationships of a given sampling period to the annual time period. This can be of greatest importance when comparing frequency distributions of pollutants to ambient air quality standards.

Based on observed APCD records of historical data (about 3 years) a graph may be prepared similar to the drawing below showing the variability of the standard geometric deviation, S_g, arithmetic mean, m, and geometric mean, m_g. It would typically look like this for primary pollutants such as CO:



The results of this study would show which time period has the closest statistical parameters to the annual average or would show approximately how far off the statistical parameters would be approximately from the annual statistical parameters based upon limited short-term sampling.

Example 25

Prediction of Worst/Worst Concentration

This is a recommended methodology for calculating the worst/worst concentration of a pollutant in the microscale area next to the roadway using Larsen's Model. Particular reference is paid to the estimation of carbon monoxide (CO) in the microscale region for transportation impact studies on air quality. This method especially lends itself to the calculation of the once-a-year expected maximum concentrations for CO for both the 1, 8 and 12 hour standard comparisons. Worst/worst conditions are considered to be defined as when the worst traffic exists with the worst meteorological conditions to produce the highest concentrations of a given pollutant---CO in this example.

The following steps should be performed for each prediction year---present, ETC, ETC+20, 'critical year'; etc:

1. Using the maximum time-distributed traffic and the worst meteorology calculate the microscale concentrations out to the desired maximum distance from the roadway. Note, to get the worst conditions for CO in the winter time at a given distance from the roadway stability categories 'D' and 'F' are normally chosen.

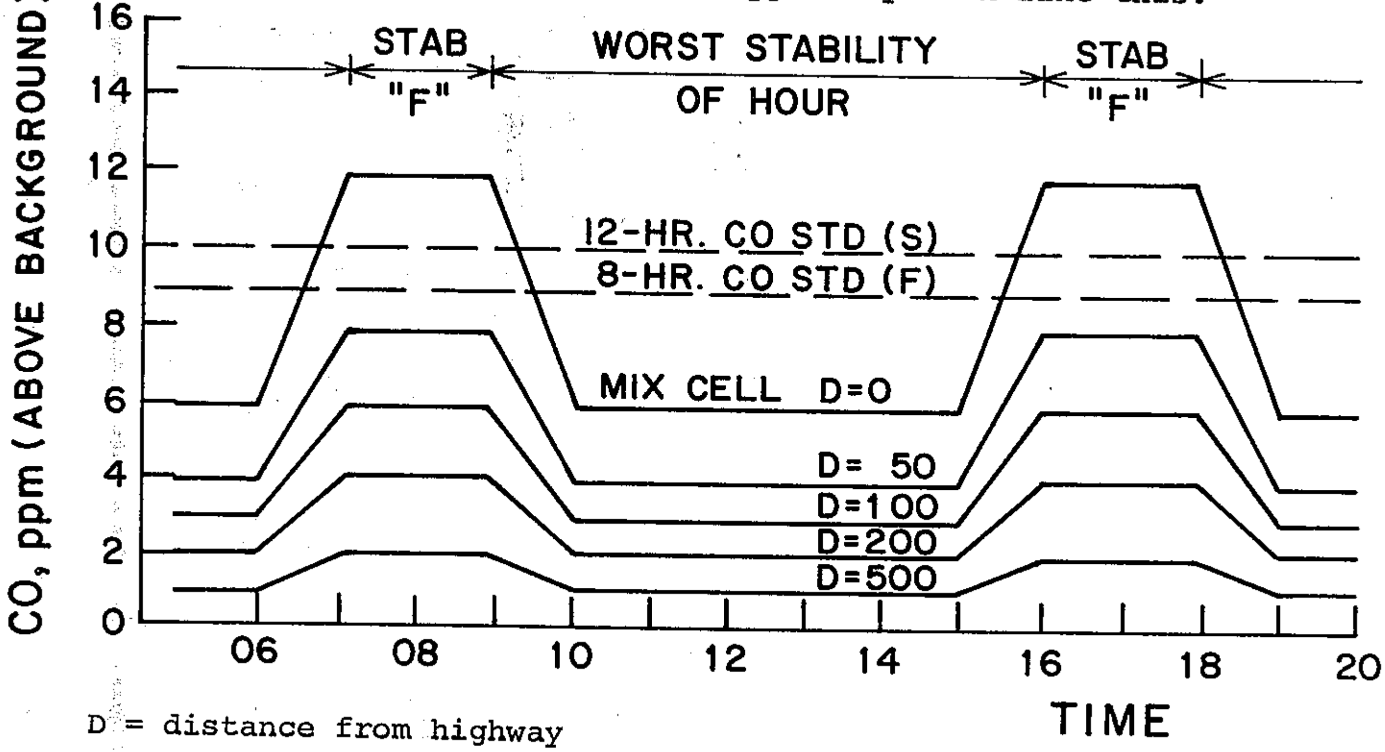
   a. Test both parallel conditions and cross-wind with phi = 130 to determine which situation will cause maximum concentrations normal to highway.

   b. Use the minimum wind speed allowable with the stability class under worst conditions, usually 2 mph.

c. Remember that under worst/worst conditions the concentrations may apply to both sides of the highway.

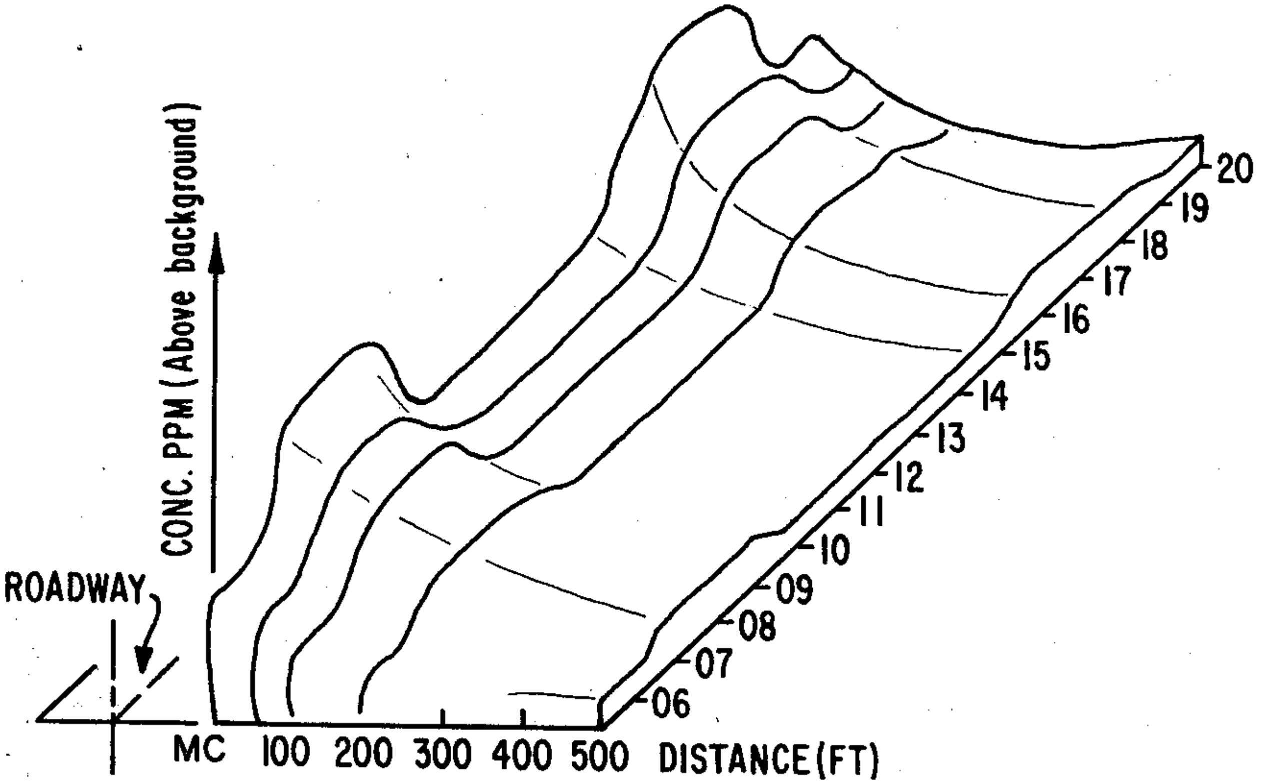
2. The results of these calculations can be used to plot "contours" of concentrations by time at a given distance from the roadway. These concentrations may be "time-averaged" or integrated (as in using a 'big' bag sampler) for 8 and 12 hour averages.

The results of the calculations would typically look like this:



D = distance from highway

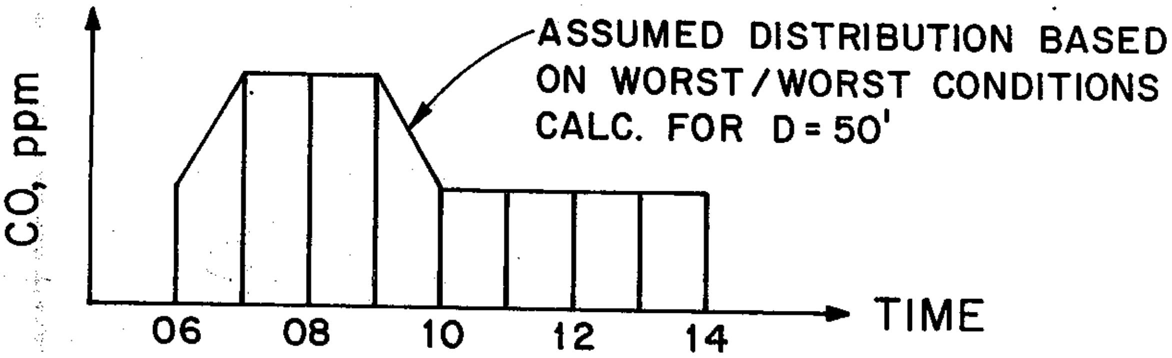
Another representation of the above is a perspective 3-dimensional view. The results would typically look like the figure below which is a representation of the concentrations shown above.



3. Calculate a running 8-hour (i.e., contiguous 8 hours) microscale average for each distance of interest from roadway, i.e., as in the figure--mixing cell, 50 ft., 100 ft., 200 ft., 300 ft., 500 ft., for the maximum 8 continuous hours of the worst/worst case. This would usually be 0600-1400, or possibly similar hours during the evening peak traffic hour, at the most stable conditions with maximum traffic.

There are at least two ways that these 8-hour microscale averages can be calculated:

a. Summing the area under the various distance contours and then dividing by eight, as in the following:



b. Or tabulating a hour-by-hour value sum of the one-hour averages and then dividing by eight.

To calculate the 12 hour CO microscale concentration the techniques would be the same; eight hour CO calculations are used here as an example. Also do present, ETC, ETC+20, and critical year also.

4. The difference between these contour levels for eight hour averages and the standard is the allowable ambient before the Federal (or State) standard is exceeded; i.e., for CO:

$C_{8\text{-Fed Std}} - C_{8\text{-micro}}$ = allowable ambient background at various distances from the roadway mixing cell.

Using the above equation and the above example:

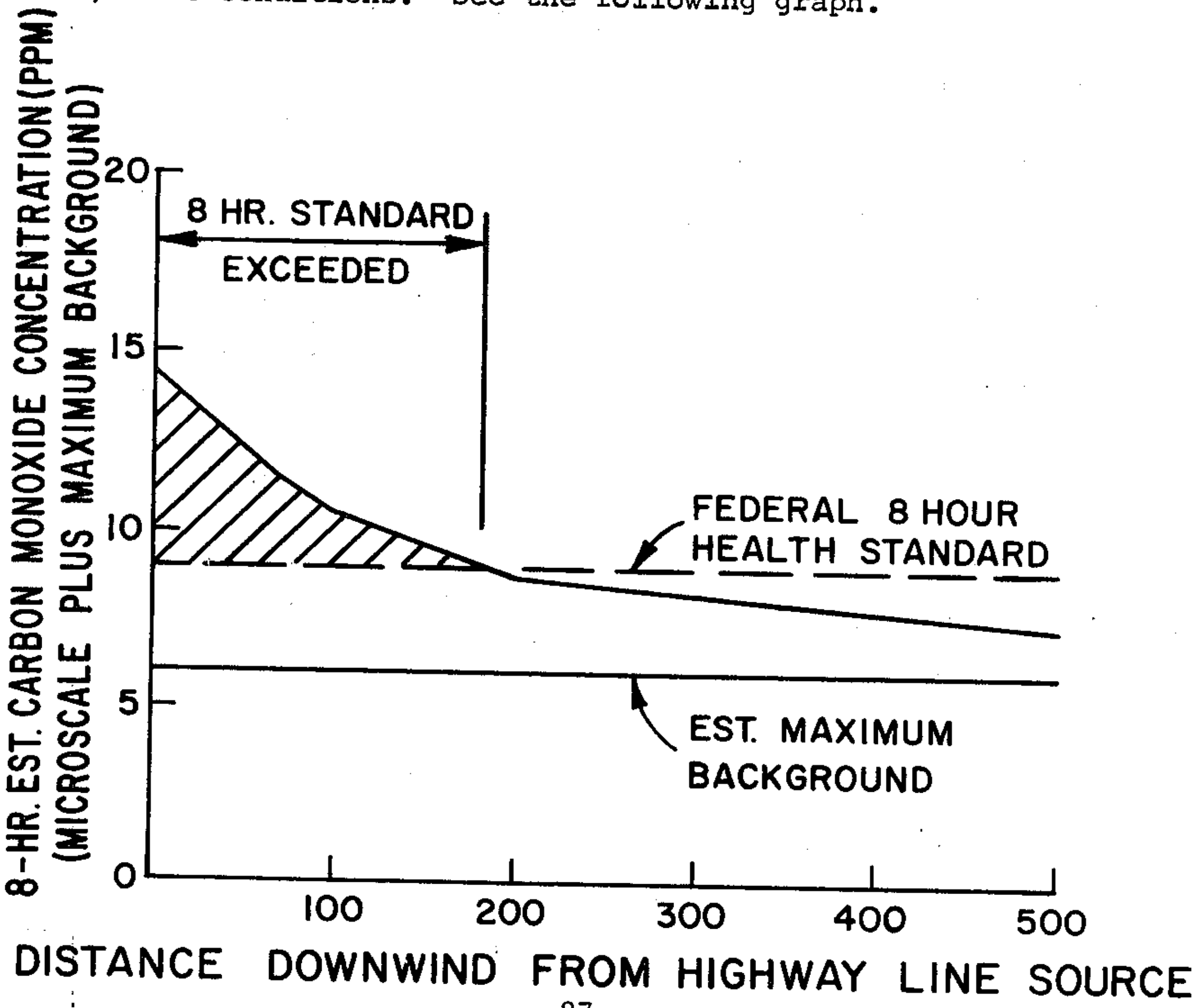
9.0 - 5.5 = 3.5 ppm allowable background under the worst/worst case before exceeding the Federal NAAQS 8-hour CO standard.

Using an example elsewhere in this set of notes that showed that the one-hour CO concentration could be represented by:

$C_1 = 2.36 * (1.38)^z$ and the eight hour average,

$C_8 = 2.42 * (1.32)^z$

where the maximum annual expected background value was 6 ppm for the 8-hour averaging time would indicate this location would exceed the 8-hour Federal standard. In fact, for the given example as shown the 8-hour CO standard would be exceeded out to somewhere between 100 and 200 feet at least once a year under worst/worst conditions. See the following graph.

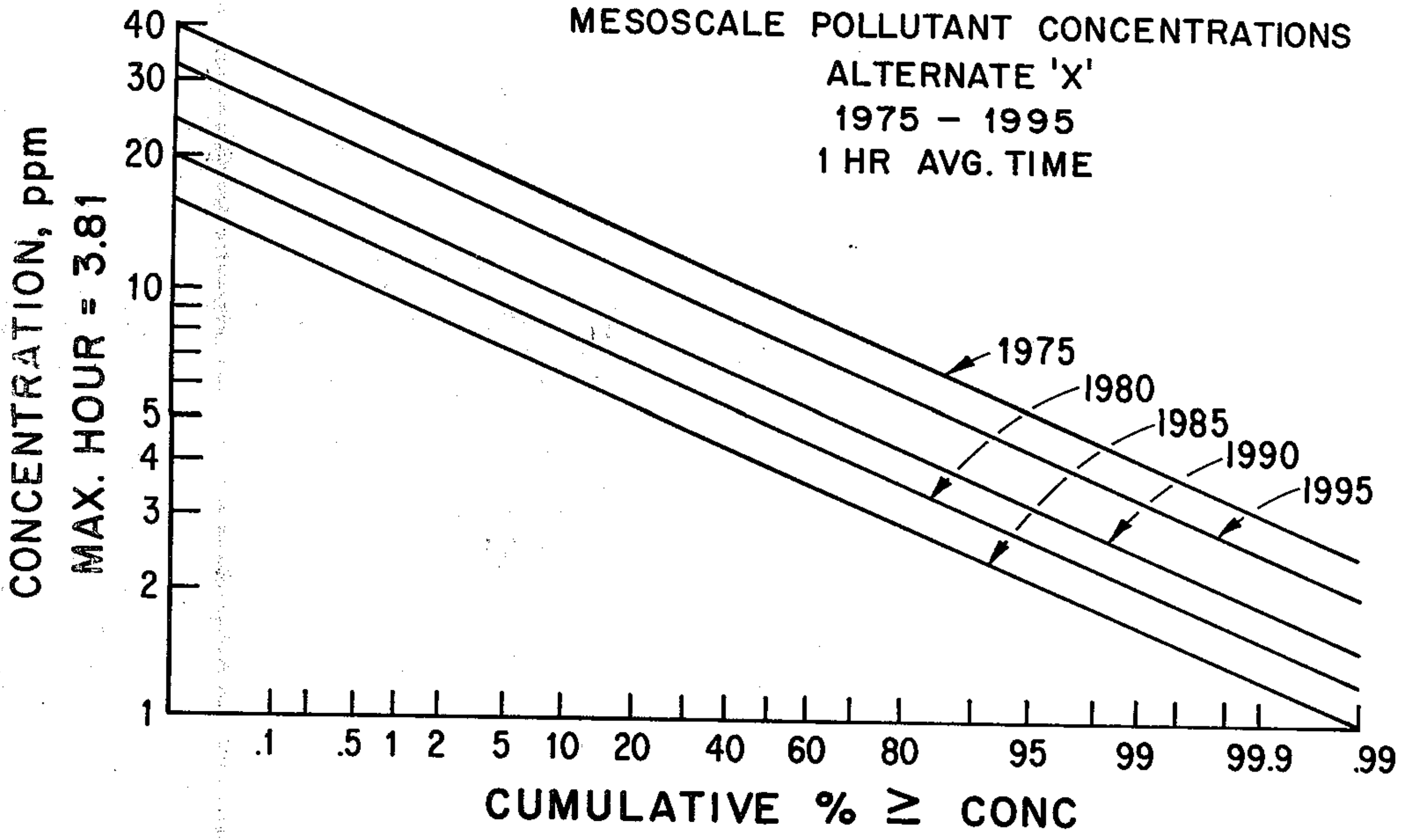


It should be emphasized that quality control of monitoring air quality data is extremely important in using Larsen's Model. All Districts should adopt a quality assurance program in field monitoring and the analysis as recommended by the Transportation Laboratory.

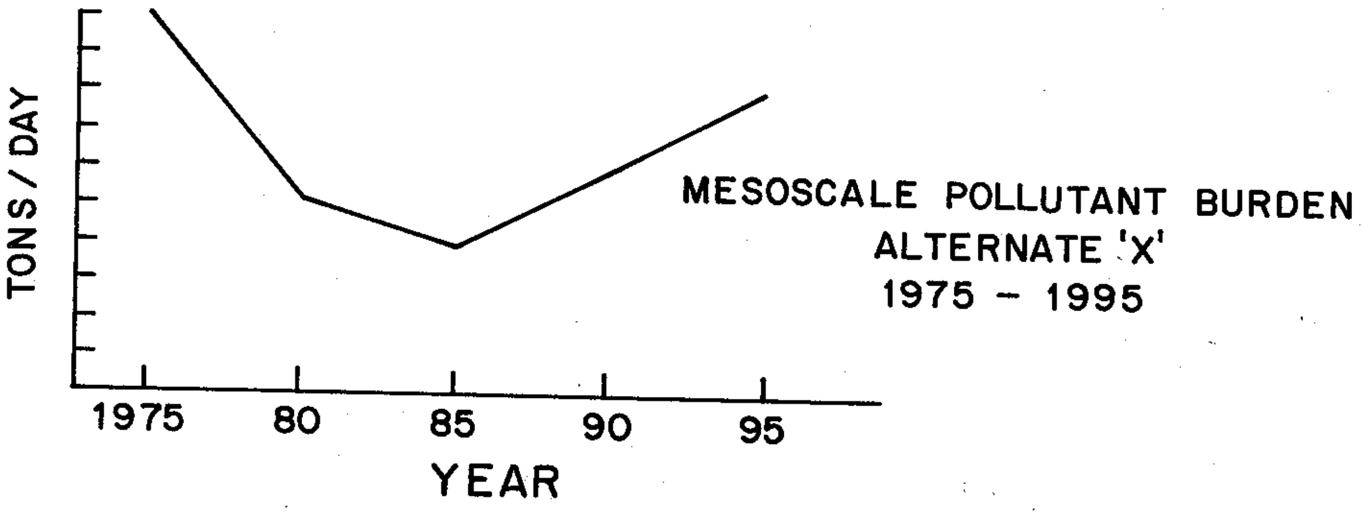
Example 26

Predicting Future Concentrations Using Tons/Day Analysis

Larsen's Model can be used in predicting future year's ambient concentrations in a "roll-back" method. This is done by assuming that the mean concentration is a function of the total pollutant burden and that the standard geometric deviation is a function of the meteorology and is relatively constant over the years. Utilizing these assumptions the following could be plotted on lognormal probability paper.



This results because of the plot of the total mesoscale pollutant burden assuming that the area is affected by the mesoscale burden only.



From the following lognormal plot of the pollutant data you should, with the aid of AP-89, be able to garner at least the following data:

YEAR	MAXIMUM CONCENTRATION 1-HOUR	MAXIMUM CONCENTRATION 8-HOUR	NO. HOURS EXCEEDING 1-HOUR STD = 35 PPM
1975	42	30	.04% = 4 hours
1980	21	15	- - -
1985	17	12	- - -
1990	26	19	- - -
1995	34	24	- - -

This is all based upon a $s_g = 1.45$ estimated.

EXAMPLE NO. 27: To Predict the Worst Pollutant Day For Urban Areas (The highest 1-hour CO concentration, both background and microscale)

1. Field Data or APCD Data (representative of highway project).

   Use Larsen's Model to estimate worst concentration assuming a log normal distribution.

*2. Use historical meteorological records associated with the worst surface stability and light winds ($\overline{U}$ = 2 mph, stability Class F, ∅ = 12.5° or parallel).

Use the CATANOVA and Friedman test to possibly eliminate the number of areas where separate microscale analyses must be performed.

Larsen's model can then be used to expand to the worst 8 hour average as long as the estimates are for an urban area and the data follows a log normal distribution.

*These meteorological conditions should be used in the line source model to predict the highest concentrations above background. The sum of the two estimates gives the highest value to address to the worst one hour health standard for CO.

EXAMPLE NO. 28: To Predict the Worst Pollutant Day for Rural Areas (1-Hour)

Approaches for Rural Areas:

1. Use Larsen's Model if it can be shown that the data is log normally distributed.

2. Use ambient air quality survey data.

3. Use highest reading at an APCD station for the last two years and use regression techniques to predict corresponding field values.

Note: Use last two years due to changes in:

1. Emission Controls

2. Traffic Patterns

3. Instrumentation

4. Site Location of APCD

4. Use historical meteorological records and then find air quality data associated with these conditions.

If data is lognormally distributed, the 8 hour concentration can be determined using Example 22. If not, use the air quality sampling data, and a moving average to determine the worst 8 hour period. If the peak value is less than 9 ppm, the highest 8 hour average will not exceed the 8 hour health standard, for the periods sampled. USE TONS/DAY ANALYSIS FOR CO TO PREDICT FUTURE BACKGROUND. Include all alternatives, including the no-build case. Show the relationship between the alternatives.

EXAMPLE NO. 29: To Predict the Worst Pollutant Day for Urban Areas (Ozone)

Use this method where ozone is considered to be a health hazard by EPA or ARB.

1. Use Larsen's Model to predict the highest one hour value.

*2. Use "rollback" technique based upon hydrocarbon emissions to reduce the predicted value for future years.

3. Include all alternatives, including the no-build case. Show the relationship between the alternatives.

*Presently the Transportation Laboratory is converting photochemical models to our computer system. When operational, this will replace the rollback technique to predict future O_3 concentrations.

EXAMPLE NO. 30 - Critical Urban Areas for NO_2

Where NO_2 is considered to be a health hazard by EPA or ARB, Larsen's Model should be used to predict both the annual average and the worst hourly average. *Use Tons/Day analysis to predict future values. Include all alternatives, including the no-build case.

1. Use Larsen's Model to predict the worst annual average and compare to the Federal health standard.

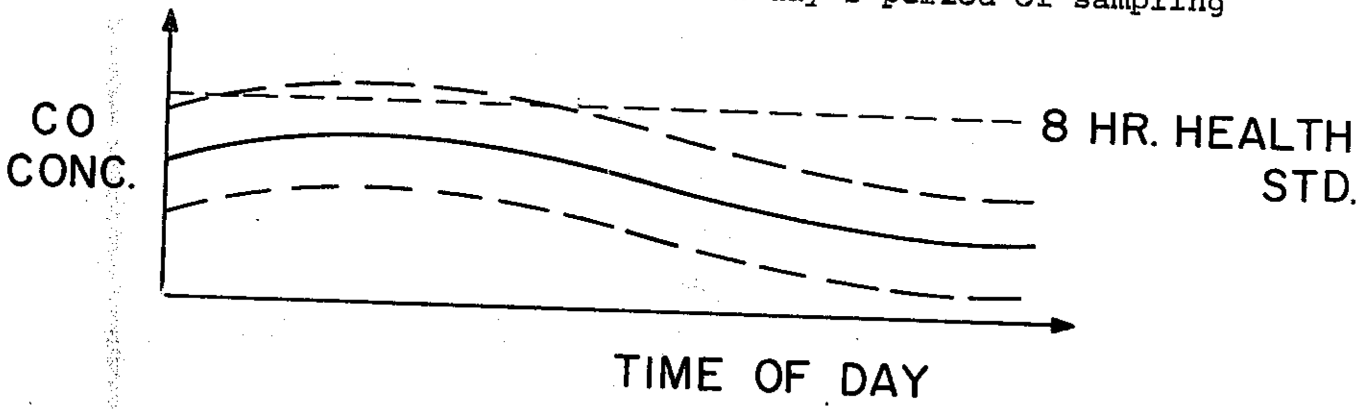
2. Use Larsen's Model to predict the worst hourly average and compare to the State standard.

3. Perform this for all alternatives including the no-build case. Show the relationship between the alternatives.

*Same as previous page. Air quality models will also predict for NO_2.

EXAMPLE NO. 31: To Predict a Typical Pollutant Day (for any season of the year)

1. Field Data - Plot the median values of the data versus time of day. See figure below. Find a day's period of sampling which is similar to the plot of the mean values. Use meteorological conditions associated with the same time period to predict the microscale contribution. Use Tons/Day method to reduce for future years. Add to the field value to obtain the overall value.



2. APCD Data - This assumes APCD data is representative of project area. Use the above procedure and consider the last two years to avoid the problems mentioned in Example 11.

3. Use the method described in the Meteorology Manual (most probable wind speed, wind direction and stability.

YEARLY CARBON MONOXIDE DISTRIBUTIONS USING REGRESSION ANALYSIS

Ambient Air Quality Standards Comparison to Estimated Yearly CO Distributions Using Regression Equations.

This method of comparison assumes that regression equations for ambient CO as a function of meteorology are known for all time periods of the day for both the high and low CO seasons. Probabilities for all "typical year" occurrences of meteorology (stability, wind direction and wind speed) are computed from a minimum of 5 years of meteorology. By use of computer, each set of typical year meteorological data is input to the proper regression equation and the dispersion model to estimate, respectively, the ambient level and the highway contributed level (at a specific distance from the highway). These levels are added and stored to the nearest whole ppm with the probability of occurrence being equal to the probability for that specific set of stability, wind direction and wind speed. When the same total concentration is computed from a different set of meteorological conditions the probabilities are summed. The resultant CO distribution is an estimation of the yearly CO distribution based on "typical-year" meteorological occurrences, for a specific year and distance downwind. This distribution has a sample size of less than 8760. A "transformation" to a sample size of 8760 through AP-89 is performed and the resultant 1 hour and 8 hour averaging time distributions are compared to air quality standards.

A more detailed description of the analysis follows:

1. Development of Ambient CO Regression Equations

   Using linear stepwise regression techniques, ambient CO prediction equations are developed for both the high and low CO seasons. Generally we will recommend the development of 6 equations per season, each representative for 2 or 3 hour time periods covering the entire day, i.e., 0200-0400, 0600-0800, 1000-1200, 1400-1600, 1800-2000, 2200-2400. (The sample size used for a CO distribution developed from 2 seasons and these time intervals would be 2,190 or 12/24 hours per day x 2/4 seasons per year x 8760 hours per year.) The CO values measured for each hour are regressed by the time intervals, with simultaneous occurrences of the following variables:

a. wind direction
b. wind speed
c. Turner's stability class (A through F)
d. inversion base height
e. your choice of any available parameter(s).

2. "Typical-Year" Meteorology

The computer program WNDROS written by M. Farrockhrooz of the Transportation Laboratory will compute frequency histograms for 16-sided wind roses and prints tables of Relative Frequency Distributions with regard to different stability classes for time periods requested by the user. This program will accept as many years of meteorological data as the user desires to include.

3. Data Analysis - Ambient CO

Using the meteorology output from WNDROS in the proper time-period ambient CO prediction equation a yearly ambient CO distribution (with known frequencies) is developed by iteration through all typical-year occurrences of meteorological variables. Inversion base height is set equal to a constant, such as, the median or mean.

The following steps a through e can now be applied to the estimated CO distribution.

a. Using equation 34 page 32 (AP-89) the standard geometric deviation is calculated from the observed maximum concentration and the arithmetic mean

$$s_g = \exp(z-(z^2-2\ln[c/m])^{0.5})$$

The z value for the observed maximum is taken from a table of frequencies.

b. The frequency for the observed maximum is calculated by equation 26 page 31.

$$f = 100\%(r-0.4)/n$$

If the observed maximum occurs more than once the rank (r) is set equal to the median rank of the value -

Number of Occurrences of Observed Maximum	Rank Used in Equation 26
1	1
2	1.5
3	2
4	2.5
n	(n/2+0.5)

For plotting purposes, the concentration is plotted to the whole ppm. This corresponds to the median of the actual pollutant interval.

c. The geometric mean is calculated using equation 23 page 30.

$$m_g = c/S_g{}^z$$

d. The maximum one-hour concentration is calculated from equation 21 page 29.

$$c_{max} = m_g \; s_g{}^z$$

e. The maximum eight-hour concentration is derived from equation 84 page 44.

$$c_{max} = c_{max\ hr} \; t^q$$

In addition, frequencies for concentrations below the maximums for 1 and 8 hour averaging times can be calculated by using equation 21 page 29 to solve for the Z values which can then be used to obtain frequencies.

$$C = M_g \; S_g{}^z \quad \text{or}$$

$$Z = \text{Logc}/(\log M_g + \log S_g)$$

4. Future Year Analysis - Ambient CO Levels

The rollback techniques are used to project ambient CO values to future levels. The CO "projection factor" is applied to the ppm values from the prediction equations. AP-89 is used on this distribution as in the previous section (3).

5. Future Year Analysis - Ambient CO + Highway Contributed CO

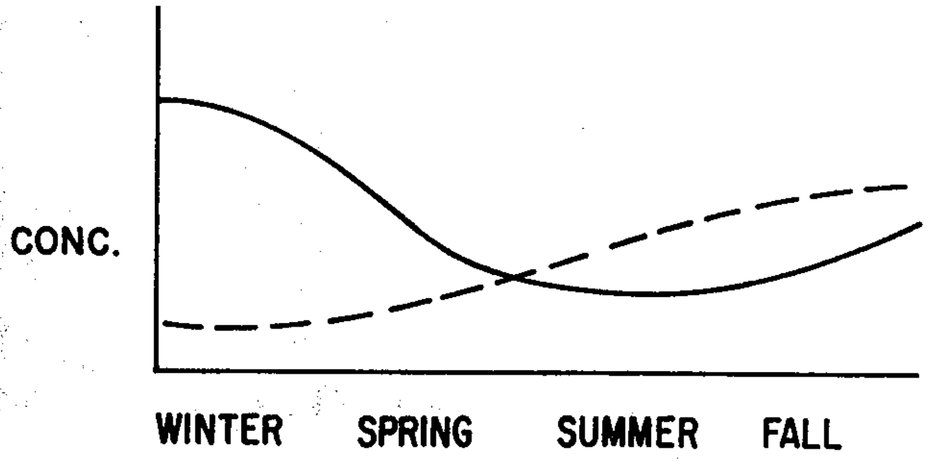
In order to obtain a yearly ambient plus highway CO distribution, a method similar to the one used in the above section is used. The meteorology output from WNDROS is used in the proper regression equation to obtain the ambient CO level for each set of meteorological parameters. The level is then reduced by the projection factor. Then the same meteorological parameters are input to the California Line Source Dispersion Model to estimate the highway CO at the various distances from the roadway. The ambient CO level is now added to the highway contributed CO level. The frequency of offurrence for this total CO level will be equal to the sum of all meteorology frequencies which lead to the same estimated total CO level.

AP-89 is applied to this estimated yearly CO distribution in the same manner as (3) also. The resultant "transformed" distribution is used in comparison to standards.

Using methodology discussed in this analysis, the yearly distributions for each averaging time can be estimated for any year at any location downwind. If the maximum value of a distribution is above the appropriate standard the number of times the standard is expected to be exceeded can be estimated since frequencies for the entire yearly distribution are known (on the assumption that a typical-year, meteorological, occurs).

STATISTICAL DESIGN OF AN AIR QUALITY SURVEY

Relative Variations of CO and O_3 vs. Seasons of the Year:



Winter Season: Generally the primary pollutants (CO, HC, NO_x) are highest and O_3 lowest.

Summer Season: Generally the primary pollutants are lowest and the secondary pollutants such as O_3 are highest.

Obtain the seasonal pollutant variation from historical records from APCD stations. This must be done before designing any air quality survey to determine the worst pollutant seasons. Refer to Figure 3.

TYPES OF AIR QUALITY SURVEYS

Purpose: To define the ambient levels for the winter (CO, HC, NO_x) season or summer (O_3)

Short Term

($t \leq 6$ months)

1. Random sampling of projects.

2. Sample every other day(s).

3. Statistical design based on daily analysis of data (applicable for one project only).

4. Generally sample as often as possible for politically or environmentally sensitive projects.

Long Term

($t > 6$ months)

1. Random sampling of projects.

2. Sample every other day(s).

3. Design based on non-parametric statistics and meteorological conditions.

4. Use of local U.S.W.B. meteorologist for synoptic weather forecasts for air pollution.

THE MAIN OBJECTIVE OF ANY AIR QUALITY SURVEY IS TO HAVE A WELL PLANNED PROGRAM OF MONITORING.

SAMPLING TECHNIQUES

A. Procedure for Randomized Block Design

1. Select priority projects.

2. Select (for each project) sampling locations based on criteria presented in "Ambient Air Quality Manual".

3. Number each site consecutively starting with one.

4. Use randomized block design to select the sampling locations to monitor for a complete day(s) using either of the following:

   a. Throw of dice
   b. Random number table
   c. Draw number from hat

BALANCED RANDOMIZED BLOCK DESIGN FOR CO & MINI-VANS

Projects or
Sampling Locations

Days of Week	1	2	3	4	5	...n	
1							Block I
2							
3							
4							
5							
⋮ n							
1							Block II
2							
3							
4							
5							
⋮ n							

B. Systematic Sampling

1. Sample every other day(s) depending on logestics.

C. Daily Sampling Periods

1. Generally 24 hours is desirable for CO. However a minimum of 12 hours per day is recommended to cover the peak AM or PM traffic hours in urban areas. Special consideration should be given to recreational traffic, weekend traffic and holiday traffic periods.

   If sampling is not randomly distributed throughout the 24-hour day, then conclusions reached using Larsen's Model are only applicable for the hours sampled.

2. To eliminate overtime, sample one day 6 AM to 1 PM and next day 1 PM to 7 PM. This assures sampling for the morning, every and off peak hours.

LEVELS OF ANALYSIS

I. It is recommended that all Districts present air quality data (CO, NO_x, HC, O_3) using the following computer programs.

1. Summary Program - Summarize data on a monthly base for each hour sampled with minimum and maximum or L.L. and U.L. values along with median.

   For special studies it may be desireable to summarize data for a few weeks time or the duration of an episode.

2. MATHISTO - Descriptive statistical program calculating means, standard deviations, ranges, histograms, cumulative frequency analysis to determine % of time that measured concentrations exceed the standard.

3. Larsen's Model to predict worst/worst 1 hour and 8 hour CO concentration.

II. For air quality studies in larger metropolitian regions or environmentally sensitive areas it is recommended that the following computer programs be used to (1) reduce field and analysis expenditures and (2) add additional crediability to the air quality reports:

A. Air Quality Surveys

1. Wilcoxin Test - correlation of sampling techniques and procedures.

2. Friedman - eliminate redundant air monitoring sites (temporal and spatial).

B. Meteorological Analysis

1. CATANOVA - Wind rose analysis
   Surface stability analysis

C. Representative year, month, season, etc.

1. Chi-square test

D. Correlation Analysis - use historical data at permanent stations if possible.

1. LINREG
2. STEPREG
3. CURFIT

REPORTING NONPARAMETRIC ANALYSES IN AIR QUALITY REPORTS

Report the following:

1. Level of significance,

2. Type of test and why the particular test was chosen.

<u>Example:</u>

Because of the small sample size and distribution for noncontinuous sampling of CO, the nonparametric Friedman test was deemed the most appropriate statistical test to analyze the data as compared to a parametric test. The test was performed at a 5% level of significance.

REPORTING REGRESSION ANALYSES IN AIR QUALITY REPORTS

Report the following:

*1. Level of significance

*2. F-value for entire equation and critical value.

3. t-value for each regression coefficient and critical value

*4 Standard error $s_{\overline{Y}}$

*5 Correlation (r) or Multiple Correlation (R) coefficient

6. Plot of Residuals - if necessary for large data set

*7 Equation $y = a+b(\mathrm{INV})+c(\frac{1}{\overline{U}})$ etc.

*Also indicate the variables considered in the analysis and tell which variables are not significant at an level.

*Should be included in all studies.

COMPUTER PROGRAMS

INTRODUCTION TO THE TENET TIMESHARE COMPUTER TERMINAL

The timeshare computer system is a user-directed computer system designed to be used with relatively few user inputs. For analyzing a large amount of data, a larger computer is required. There are three operating modes in the tenet system, "executive", "basic", and "editor".

When the computer is turned on (generally one switch for the printer and another one to tie to the computer), the operator is told to LOGIN. This means to type in the account number and name that has been assigned to the user for his use (the account number and name must be separated by a semicolon).

If a password is used in order to eliminate unauthorized use of the account, the computer prints PASSWORD?. The user then types in the password, which is automatically not printed on the paper for security reasons.

When this has been done the computer automatically is in the "executive" mode. The computer signals that it is ready to accept anformation by printing the executive prompt (-). In this mode the user is able to create files, read paper tapes into and out of the computer, delete files, append one file to another, copy files from the memory of the computer to the printer, "shift" to another operating mode, and many more operations.

When in the executive mode (-) and EDIT is typed, the computer sHifts into the "editor" mode. The computer then prints the edit prompt (@). As the name implies, the user is able to modify or edit files in this mode. First the command LOAD '(filename)' must be given. After the errors are all corrected and an @ sign is printed by the computer, type SAVE OLD '(filename)'. Then type Q CLEAR to return to the executive mode. Many additional operations can also be performed in the edit mode.

To go from the executive mode to the basic mode, type BASI. The computer will print the basic prompt (>) when this move has been completed. The computer is now ready to use the computer program that the user selects. This is done by typing LINK (occasionally LOAD is used) followed by the name of the computer program, enclosed within quotation marks.

An example would be the following:

```
LINK '5;LSTAT;LINREG'
```

In this example the name of the computer program is 5;LSTAT;LINREG. It must always be within quotation marks. When the computer program is finished the computer will once again print the basic prompt (>). At this time the user can type RUN to run the program again, another LINK command to use another computer program, or QUIT to return to the executive mode.

In the example which follows, the computer program was interrupted by hitting the escape combination. This combination can vary from terminal to terminal but the most common combinations are:

"Shift and]"
"Control and ["
"Control and shift and ."

The computer prints a (#) sign and prints information concerning what the computer was doing when it was interrupted.

When the user is finished with his work and wishes to leave the computer system, he must be in the executive mode. When the executive prompt (-) is printed, the user types LOGO. The computer will print the date, followed by how much processing time was used (CPU time), how much terminal time was used, and

```
           TENET 210 TIME-SHARING SYSTEM  0905  05/13/74 4D

-LOGIN  452;STAR
PASSWORD?

-EDIT
@LOAD 'REG'
   20  RECORDS AFFECTED
@LIST .5:5
       .50 COMPARING APCD SAMPLING WITH BAG SAMPLING,19
     1.00    3   3
     2.00    5   4
     3.00    5   4
     4.00    5   5
     5.00    6   5
@SAVE OLD 'REG'
   20  RECORDS AFFECTED
@Q CLEAR
-LINK '5;LSTAT;LINREG'
EH?
-BASI
>LINK '5;LSTAT;LINREG'

              ***** MODIFIED 12/6/72 *****

   INPUT DATA FILENAME OR "EXP" FOR PROGRAM EXPLANATION?#

INTERRUPT DURING LINE  150
>QUIT
-LOGO
 0907  05/13/74
CPU MINS -  0.014
TERMINAL MINS -   1.70
FILE MODULES - 21
----------------
```

how many memory spaces are being used by the particular account number and name that was being used by the computer user. The computer will then print a horizontal line. At that time, the user turns off the 2 (typically) switches that he turned on to start the computer. The following is an example printout of the above instructions.

There are several methods of inputing the data into the memory of the computer:

1. Prepare a punched paper tape and read the tape into the computer.

2. Type the data into the computer while in the Edit mode, or

3. Type the data into the computer while in the Executive mode.

The third method will be briefly described. After the LOGIN operation has been completed, type:

```
COPY TEL TO (filename)
```

The computer printer will return to the left side after the carriage return is hit, but it will not print the executive prompt (-). Type in the data as shown on the example which follows. When the data has all been entered, hit "control D". The computer will once again print the executive prompt. At this time, another data file could be created, or any other operation could be performed.

APPLICATIONS OF COMPUTER PROGRAMS

The general problem which is presented as Figure 1 will be used to demonstrate the procedures necessary in the use of the statistical methods which have been described in the first part of this course. The necessary steps are the following:

1. Determine whether the data from mechanical weather stations are different from the data obtained at other meterological sites (including other mechanical weather stations).

2. Show that the sampling techniques that you used are compatible with those employed by other agencies.

3. Determine whether there are any redundant air quality sampling locations.

4. Determine whether the year in which you sampled could be considered a "typical" year.

5. Use historical meteorological and/or ambient air quality data to estimate the concentrations of pollutants for the time periods that you did not sample.

6. Use Larsen's Model to predict worst pollutant concentrations.

7. Use statistics to show what percent of the time the air quality standards were exceeded.

8. Use summary program to present the data in a neat, tabular form.

CATEGORICAL ANALYSIS OF VARIANCE

This procedure, commonly called "Catanova", is used to determine whether any two or more sets of meteorological data are statistically identical. If this is the case, then it will eliminate redundant data as inputs into air quality models or it may not be necesssary to measure wind speed and direction at all the locations. The program that performs this comparison is called "5;LSTAT;CATANOVA". To perform this test, data must be obtained for the same time periods at all the locations of interest.

A typical use would be to compare the frequency of occurrence of the observations from one mechanical weather station with the frequency of occurrence of the observations from another mechanical weather station. For a complete comparison, this procedure would be performed for morning, midday and evening data for all seasons of the year. This data can be obtained from the summaries of the "STAR2", "WNDROS", or "WIND2" computer printouts for the time periods and months of interest and at each site. It is necessary to obtain current data from any historical site.

Page 137 of the Meteorology manual will be used to show where the data is lifted from the computer summary for the Hayward Airport.

The following table contains the data required to test whether there is a significant difference in the occurrence of wind speed and direction for the Airport, at Location #1 and at Location #3 as depicted on Figure 1.

WEATHER STATION NAME: HAYWARD AIRPORT LOCATION: HAYWARD CAL
YEARS OF RECORD: 66-67-68-69-70 MONTHS: SEPT.-OCT.
HOURS OF DAY: 0700-0800-0900 COMBINED
PROJECT ID: 19701 761650 USER: AR DISTRICT: 19

STABILITY CLASS DESIGNATION 9

IN THE FOLLOWING TABLE THE CALMS ARE DISTRIBUTED....

.......FREQUENCY DISTRIBUTION TABLE.......

DIRECTION	VELOCITY,MPH											
	0-3	4-7	8-12	13-18	19-24	25-31	32-38	39-46	47	TOT	AVE.	%TO
N	2	24	1	0	0	2	0	0	0	30	6.9	3.2
NNE	4	18	5	0	0	1	0	0	0	28	6.5	3.1
NE	9	36	15	1	0	0	0	0	0	61	6.2	6.6
ENE	4	101	45	2	0	0	0	0	0	152	6.8	16.6
E	11	60	15	0	0	0	0	0	0	86	5.8	9.4
ESE	7	40	14	4	0	0	0	0	0	65	6.7	7.1
SE	4	42	22	7	1	1	0	0	0	78	7.9	8.5
SSE	2	24	16	3	2	1	0	0	0	49	8.5	5.3
S	0	20	3	0	0	0	0	0	0	23	6.1	2.5
SSW	2	11	7	2	0	0	0	0	0	22	7.4	2.4
SW	4	12	9	6	1	0	0	0	0	32	8.6	3.5
WSW	9	22	14	9	1	1	0	0	0	56	8.3	6.1
W	4	56	21	13	2	1	0	0	0	97	8.2	10.6
WNW	2	33	14	2	0	0	0	0	0	52	6.9	5.6
NW	4	33	12	2	1	0	0	0	0	53	6.9	5.8
NNW	4	20	5	2	1	0	0	0	0	32	6.8	3.6
CALM	346	0	0	0	0	0	0	0	0	346	.0	.0
TOT	76	552	218	53	9	7	0	0	0	915	.0	.0

TOTAL NO. OF OBSERVATIONS = 915
OCCURRENCE WITHIN THIS STABILITY CLASS = 915

..........RELATIVE FREQUENCY DISTRIBUTION..................

RELATIVE FREQUENCY OF OCCURRENCE OF ALL STABILTTY CLASS= 100.0 PCT

DIRECTION	VELOCITY,MPH									
	0-3	4-8	8-12	13-18	19-24	25-31	32-38	39-46	47	%TO
N	.24	2.68	.11	.00	.00	.22	.00	.00	.00	3.25
NNE	.49	1.95	.55	.00	.00	.11	.00	.00	.00	3.09
NE	.97	3.89	1.64	.11	.00	.00	.00	.00	.00	6.62
ENE	.49	10.95	4.92	.22	.00	.00	.00	.00	.00	16.58
E	1.22	6.57	1.64	.00	.00	.00	.00	.00	.00	9.43
ESE	.73	4.38	1.53	.44	.00	.00	.00	.00	.00	7.08
SE	.49	4.62	2.40	.77	.11	.11	.00	.00	.00	8.50
SSE	.24	2.68	1.75	.33	.22	.11	.00	.00	.00	5.32
S	.00	2.19	.33	.00	.00	.00	.00	.00	.00	2.52
SSW	.24	1.22	.77	.22	.00	.00	.00	.00	.00	2.4
SW	.49	1.22	.98	.66	.11	.00	.00	.00	.00	3.46
WSW	.97	2.43	1.53	.98	.11	.11	.00	.00	.00	6.14
W	.49	6.08	2.30	1.42	.22	.11	.00	.00	.00	10.62
WNW	.24	3.65	1.53	.22	.00	.00	.00	.00	.00	5.64
NW	.49	3.65	1.31	.22	.11	.00	.00	.00	.00	5.78
NNW	.49	2.19	.55	.22	.11	.00	.00	.00	.00	3.5

	AIRPORT	LOCATION #1	LOCATION #3
N	7	6	3
NNE	4	5	2
NE	12	14	16
ENE	30	30	26
E	16	20	12
ESE	15	12	18
SE	18	16	18
SSE	10	13	12
S	4	4	2
SSW	3	3	3
SW	6	7	7
WSW	10	10	13
W	22	20	23
WNW	10	8	12
NW	11	10	11
NNW	6	6	6
*	*	--	--
TOTAL	184	184	184

*Note: Do not include Calms if they were distributed.

	HAYWARD AIRPORT	LOCATION #1	LOCATION #2
0-3	18	14	16
4-7	108	120	101
8-12	46	40	52
13-18	10	9	12
19-24	2	1	2
> 24	0	0	1
TOTAL	184	184	184

The following pages are the actual computer data files and the computer run of '5;LSTAT;CATANOVA'. The data filenames were chosen for convenience. All instructions given to the computer by the operator are underlined. After the underlined instructions are typed, hit the carriage return.

```
-COPY CAT%DIRE TO TEL TEXT


  7   6   3
  4   5   2
 12  14  16
 30  30  26
 16  20  12
 15  12  18
 18  16  18
 10  13  12
  4   4   2
  3   3   3
  6   7   7
 10  10  13
 22  20  23
 10   8  12
 11  10  11
  6   6   6
-
```

```
 COPY CAT%SPED TO TEL TEXT
  18  14  16
 108 120 101
  46  40  52
  10   9  12
   2   1   2
   0   1   2
-
```

```
ANOVA FOR CATEGORICAL DATA

COMPARE AIRPORT & LOCATION 1 & LOCATION 3--DIRECTIONS


DATA:

               1         2         3
-------------------------------------


  1--          7         6         3
  2--          4         5         2
  3--         12        14        16
  4--         30        30        26
  5--         16        20        12
  6--         15        12        18
  7--         18        16        18
  8--         10        13        12
  9--          4         4         2
 10--          3         3         3

 11--          6         7         7
 12--         10        10        13
 13--         22        20        23
 14--         10         8        12
 15--         11        10        11
 16--          6         6         6


          CATANOVA TABLE

      SOURCE              SS

BETWEEN GROUPS          .30435     % EXPLAINED =          .12053
WITHIN  GROUPS       252.20108     CHI-SQUARE  =         9.96203
                                   DEG.FREEDOM =       30
           TOTAL     252.50543     PROBABILITY =          .99978

        DO NOT REJECT THE HYPOTHESIS THAT THE 3 GROUP POPULATIONS
ARE THE SAME FOR THE 16 CATEGORIES.  (SIGNIFICANCE LEVEL = .050)
```

```
RUN

        DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?CAT%SPED

        JOB TITLE?COMPARE AIRPORT & LOCATION 1 & LOCATION 3--WIND SPEEDS

        NUMBER OF:
           ROWS (CATEGORIES)     ?6
           COLUMNS (GROUPS)      ?3

        SIGNIFICANCE LEVEL?.05
```

```
COPY %REGXFER TO TEL TEXT


1000 OPEN '$REG',1,IO,BINARY,RANDOM,OLD
1010 DOUBLE TYP
1020 INPUT FROM 1 AT 1:HED$,F$,TYP
1030 PRINT;'ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)':
1040 INPUT AN$
1050 IF AN$='NO' THEN 1170
1055 IF TREC(1)<5 THEN 1070
1060 ERASE FILE 1 FROM 5 TO TREC(1)
1070 ON TYP GOTO 1100, 1120
1080 REM
1090 REM
1100 REM
1110 REM
1120 LINK '5;LSTAT;STP2'
1130 REM
1140 REM
1150 REM
1160 REM
1170 PRINT;"INPUT NEW DATA FILENAME OR 'STOP'":
1180 INPUT F$
1190 ERASE FILE 1 FROM 1 TO TREC(1)
1200 IF F$='STOP' THEN 1390
1210 PRINT;'NEW JOB TITLE':
1220 INPUT HED$
1230 ON TYP GOTO 1240, 1280
1240 REM
1250 REM
1260 REM
1270 REM
1280 HED$="'"+CHAR(12)+"'//6B'STEPWISE MULTIPLE LINEAR REGRESSION'"
      +"///6B'"+HED$+"'///"
1290 GOTO 1340
1300 REM
1310 REM
1320 REM
1330 REM
1340 PRINT ON 1 AT 1:HED$,F$,TYP
1350 LINK '5;LSTAT;CORRE'
1360 REM
1370 REM
1380 REM
1390 PRINT
1400 PRINT;;"GOOD-BYE.  BE SURE TO DELETE FILE '$REG' BEFORE LEAVING"
1410 PRINT;'THE SYSTEM.'
-
```

The data is copied out of the computer memory and onto the printer by the command COPY CAT%DIRE TO TEL TEXT which is typed into the computer when it is in the Executive mode. When the computer completes the data file, the Executive prompt is printed (-). At this time, repeat the above command with CAT%SPED replacing CAT%DIRE to obtain the printout of the next data file.

At the completion of the second data file, the computer is ready to exercise the statistical program (or do any operation, really). Usually it is convenient to hit the "Home" button on the computer to start the program at the top of a page. The next instruction necessary is to type BASI to place the computer in the Basic mode. When the computer has entered the Basic mode, the computer prints the Basic prompt (>). At this time type LINK '5;LSTAT;CATANOVA' to start the computer program. If you wish to obtain program information, type EXP. If not, type the name of the data file In this example the data file name is CAT%DIRE. Any descriptive job title can be used. The number of rows is requested next. This can be counted from the data file printout. The number of columns is entered next, also obtainable from the data file printout. Input .05 when asked for the significance level.

The computer then automatically advances the page and starts running the program. It prints the title of the statistical test, followed by the job title selected by the operator. Next, the data is printed so that the operator can be sure that the data is correct.

The "total SS" (sums of squares) is summed from the total for each row (the first row's value is generated from 16 in this example). The "within groups SS" (sums of squares) is generated from the square of each value in the data matrix, divided by the number of observations in the column in which the value is found. The "between groups SS" is obtained by subtracting the "within SS" from the "total SS".

The Chi-square is calculated from the data and compared against a table value at the requested level of significance. If the calculated chi-square is smaller than the statistical value of chi-square we can say that there is no statistical difference in the frequency of occurrence of the data. Accept the null hypothesis. In this example, at a level of significance of .05 and with 30 degrees of freedom the calculated chi-square, 9.96 is smaller than the statistical value of chi-square, 43.77. Also note that the probability is greater than .05 in our example. Thus there are two ways of reaching the same conclusion.

If the results from the CATANOVA test indicate that there is a significant difference in two sets of data, say for wind speed at the Airport and at Location #1 for example, it is possible that regression techniques could be employed to generate a statistical relationship between the sites. This relationship could be used to quantify the differences between the sites. Note, however, that the CATANOVA test requires frequency of occurance of data, while the regression test requires the actual data. Refer to the section on regression for more information.

Special precaution is required if regression techniques are employed for this analysis. First, the observations for the same hour must be paired. Second, since the starting threshhold of the meteorological instruments may not be the same, it is possible to have calms reported for one instrument's set of data but not for the other one. To avoid the necessity of assigning a numerical value to a reported calm, remove all data pairs which contain a calm.

WILCOXON MATCHED SIGN TEST

Another important step is to determine if the sampling scheme that we employ (bag sampling) yields the same value as the concentrations measured by local agencies such as APCD's. This adds credibility to the validity of our measurements and assures local officials that they would have obtained the same values as ourselves had they been sampling at our sampling locations.

There are two methods that can be employed to solve this problem. If there are enough data pairs regression techniques could be employed to correlate our measurements to other measurements. A detailed discussion of the use of regression is found in another part of this set of notes. As we usually have only 8-12 data pairs available for this analysis, or if there is not a very great spread in the data, usually the Wilcoxon Matched Sign Test is used. The computer name for this test is 5;LSTAT;MPAIR. The following is a sample of the type of data that is required to perform this test:

TIME	APCD	BAG SAMPLING
0800	1	1
0900	3	2
1000	3	4
1100	5	4
1200	6	6
1300	4	5
1400	3	3
1500	2	2
1600	1	1
1700	0	1

The following is a sample of three computer runs which illustrate the use of this test.

```
-COPY WILCOX1 TO TEL TEXT


  1  1
  3  2
  3  4
  5  4
  7  6
  5  4
  3  3
  2  2
  1  1
  0  1
-
```

```
-BASI
>LINK '5;LSTAT;MPAIR'

        DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?WILCOX1
        JOB TITLE?COMPARE APCD SAMPLING TO BAG SAMPLING

        NUMBER OF:
           ROWS (OBSERVATIONS)?10
           COLUMNS (DISTRIBUTIONS)?2
```

```
WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

COMPARE APCD SAMPLING TO BAG SAMPLING


OF THE 10 PAIRED OBSERVATIONS, 4 HAD POSITIVE
DIFFERENCES AND 2 HAD NEGATIVE DIFFERENCES.

SIGNIFICANCE LEVEL (2-TAILED)
   1 = .10
   2 = .05
INPUT INTEGER SELECTION?2

THE VALUE OF T FOR N = 6 IS 7


CRITICAL VALUE = 1



DO NOT REJECT THE HYPOTHESIS THAT THE TWO DISTRIBUTIONS
ARE FROM THE SAME POPULATION. (SIGNIFICANCE LEVEL = .050)



    ANOTHER ANALYSIS USING THE SAME DATA?NO
    ANOTHER PROBLEM IN FILE WILCOX1?NO
>
```

The data is put into the computer in the same manner as for the CATANOVA test. Once the data file has been verified to contain the correct data, type BASI. When the "basic" prompt is printed, (>) type LINK '5;LSTAT;MPAIR'. The computer will then ask for the data filename or EXP if program information is requested. In this example the file name WILCOX1 was entered. The computer will ask for a job title. Any descriptive title can be used. The number of rows are entered and then the number of columns are entered.

The computer prints out the name of the test and the job title selected by the user. It then prints the results. In this example there were 4 instances where the first value was larger than the second value and 2 instances where the second value was larger than the first value. There were 4 instances where there were no differences in the observations for a total of 10 sample pairs. The operator has the choice of testing at a 5 or 10 percent level of significance. In this example, 2 was entered to test at the 5 percent level. It is recommended that a 5 percent level be used.

The computer determines the values N and T. In this example, T = 7 and N = 6. The critical statistic value is printed and it is equal to 1 in this example. If the "T" value is greater than the critical value, then we can say that there is no significant difference between the methods of sampling.

The data file WILCOX is the same as WILCOX1 except that observation number 5 is a tie and observation number 6 was reversed. In this case, there are not enough data pairs with differences to test at the 5 percent level of significance and the test must be performed

```
-COPY WILCOX TO TEL TEXT



    1     1
    3     2
    3     4
    5     4
    6     5
    4     5
    3     3
    2     2
    1     1
    0     1
-
```

```
RUN

     DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?WILCOX
     JOB TITLE?COMPARE APCD SAMPLING TO BAG SAMPLING

     NUMBER OF:
        ROWS (OBSERVATIONS)?10
        COLUMNS (DISTRIBUTIONS)?2
```

```
WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

COMPARE APCD SAMPLING TO BAG SAMPLING


OF THE 10 PAIRED OBSERVATIONS, 2 HAD POSITIVE
DIFFERENCES AND 3 HAD NEGATIVE DIFFERENCES.

SIGNIFICANCE LEVEL (2-TAILED)
   1 = .10
INPUT INTEGER SELECTION?1

THE VALUE OF T FOR N = 5 IS 6


CRITICAL VALUE = 1


DO NOT REJECT THE HYPOTHESIS THAT THE TWO DISTRIBUTIONS
ARE FROM THE SAME POPULATION. (SIGNIFICANCE LEVEL = .100)


     ANOTHER ANALYSIS USING THE SAME DATA?NO
     ANOTHER PROBLEM IN FILE WILCOX?NO
>
```

at the 10 percent level. An alternate solution would be to use regression techniques if there are enough data pairs and the data has a reasonable range.

Data file WILCOX2 is the same as WILCOX1 except that observation numbers 3 and 5 are made into ties and observation number 6 was reversed. In this example there were not enough pairs with differences to perform the test. In this case, try regression techniques or collect more data and use the WILCOXIN test again.

Note that in the three examples just described, the difference between the data pairs is zero or one. The values are all within the accuracy of the instruments. Since this is the case, one could say that there is no difference between sampling techniques even if the statistical analysis were to indicate otherwise.

```
-COPY WILCOX2 TO TEL TEXT


   1    1
   3    2
   4    4
   5    4
   6    6
   4    5
   3    3
   2    2
   1    1
   0    1
-
```

```
BASI
>LINK '5;LSTAT;MPAIR'

      DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?WILCOX2
      JOB TITLE?COMPARE APCD SAMPLING TO BAG SAMPLING

      NUMBER OF:
         ROWS (OBSERVATIONS)?10
         COLUMNS (DISTRIBUTIONS)?2
```

```
WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

COMPARE APCD SAMPLING TO BAG SAMPLING


OF THE 10 PAIRED OBSERVATIONS, 2 HAD POSITIVE
DIFFERENCES AND 2 HAD NEGATIVE DIFFERENCES.

UNABLE TO TEST 4 PAIRS WITH DIFFERENCES - 5 MINIMUM.
```

GENERALIZED FRIEDMAN TEST

The Friedman Test can be used to determine whether the air quality sampling sites that have been selected are all necessary from a statistics point of view. Regardless of the statistical outcome, it may be desireable to retain sampling sites which are statistically redundant but are important for other reasons such as politically or environmental sensitive areas.

Data Grouping For Friedman 2-Way ANOVA

		SITE							
DATE	HOUR	1	2	3	4	5	6	7	8
5-1-73	0600								
"	0700								
"	0800								
"	0900								
5-2-73	0600								
"	0700								
"	0800								
"	0900								
5-3-73	0600								
"	0700								
"	0800								
"	0900								
5-4-73	0600								
"	0700								
"	0800								
"	0900								
5-5-73	0600								
"	0700								
"	0800								
"	0900								
ETC	ETC								

The Friedman test will be of value both during sampling and after all sampling is complete. When used during the sampling period, it may enable the user to reduce the number of air quality sampling locations. When used after all sampling is completed at a site, it will tell if too much sampling was performed and will help determine how many sampling locations should be considered for future projects with a similar layout. It may also reduce the amount of analysis in determining what background levels to use for the project area.

Prepare a data file in the following format:

		SITE							
DATE	HOUR	1	2	3	4	5	6	7	8
5-1-73	0600								
"	0700								
"	0800								
"	0900								
5-2-73	0600								
	0700								
	0800								
	0900								
5-3-73	0600								
	0700								
	0800								
	0900								
5-4-73	0600								
	0700								
	0800								
	0900								
5-5-73	0600								
	0700								
	0800								
	0900								
ETC	ETC								

Each column contains the data for any one site. The data is in chronological order. The computer will test the data in the data file to see if the various sampling sites are statistically alike for the hours in the data file.

The program that is used to perform this test is 5;LSTAT;FRIED. After the basic prompt, (>) is printed by the computer, type LINK '5;LSTAT;FRIED'. The computer will ask for the data file name or EXP (if program information is desired). The computer will then ask for a job title. Any descriptive job title may be used.

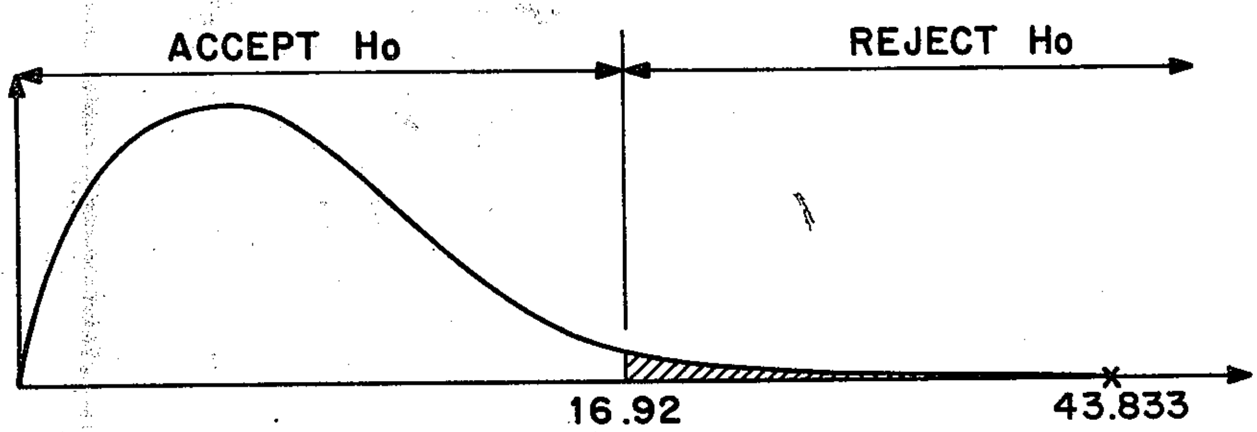
Next, the computer will request the number of observations per replication. Input the number of hours per day that data were collected. When asked the number of columns (treatments) enter the number of sites in the data file. When asked for the number of replications, enter the number of sampling days in the data file.

The computer will then advance to a new page and print the test title followed by the job title. It will then print the data, separated by replications (days). Check this data against the data file to insure that the program has been initiated correctly.

The page will be advanced again and the test title and job title again will be printed. Next the number of columns (sites) that the user wishes to compare will be asked for. For the first analysis, input the total number of sites in the data file. The computer will then replace the data with the rank of the data. When assigning ranks, it is assumed that for any given hour, the concentrations at all the sites in the analysis and for all the replications have an equal probability of occurrence. Thus, the concentration which is the lowest for each one hour block is given the rank of one. See the blocks on page 141.

The computer will once again move to a new page and print the column rank totals. These totals are for each site in the analysis. If the column rank totals are statistically alike, then there is no difference between sites. The significance level (α level) at which the test is to be performed is requested next. Enter .05 (always use decimal point).

The calculated chi-squre is printed along with the degrees of freedom and the probability. The computer prints the conclusion statement. This can be checked by referring to the table of critical values on page 286. With 9 degrees of freedom, the statistical value of chi-squre is 16.92. Since the calculated chi-square, 43.833 is greater, reject the hypothesis that the data are the same. See the following sketch.



When asked if another analysis using the same data is requested, type YES at this time. It appears that sites 4, 6 and 9 have very similar values. To test these three sites, input 3 for the number of columns in the analysis. The computer now asks the user to input 3 columns in ascending order. The ranks will be reassigned, considering only the sites selected by the user. The column rank totals are printed for these sites. The calculated chi-square and degrees of freedom are printed. As can be seen by checking the table of critical values, the computer printed the correct decision.

The test was rerun again, testing sites 1, 7, 8 and 10. These sites were also similar statistically. Other sites could now be tested to see if a further reduction in the number of sites could be made.

```
COPY %FRIED TO TEL TEXT
  5   5   3   4   5   3   6   4   6  10
  4   5   3   5   5   4   5   4  11  11
  6   5   2   7   5   3   5   6  10  10
 12   8   2   4   5   4   8   9   6   8
 13  14   3   4   6   5   9  14  10   9
  7  10   3   3   5   4   8  13   9   8
  8   5   4   7   4   8   7   5   5   8
 12  12   6   8   6  10   7   9  10  12
 13   8   8   8   8   7   7   7   8   9
  7   8   3   4   6   7  10  10   5  11
 15  13   6   4   7   9  14  14   8  13
 12  10   6   7   8   5   6  14  11   9
 13   8   3   6   7   8  10   8   5  11
 20  11   4  13  10  10  16  14  10  13
 13  13   6   8  10   8  10  10   9   9
 13   6   5   8   6   9  10  10   8  14
 17   8   9  10   8  14  12  14  11  21
 14  10   8  10  11   8   7   8  10  13
  8   8   3   4   5   5  12  10   7   9
 17  14   3   5   8   6  13  14   8  11
  8  11   4   5   5   6   9  18   8   8
 20  12   8  15  14  14  15  13   8  18
 21  16  16  13  20  16  18  23  11  19
 14  13  10  11  10  14  14  12  14  13
 21  18   9  14  21  12  14  16   7  13
 33  25  14  22  25  14  16  22  11  16
 16  16  10  13  20  15  19  16  11  15
 30  13  15  22  24  17  24  19  16  24
 34  10  24  29  27  30  30  21  21  20
 27   8  21  25  22  22  21  22  17  18
 13  12  10  11  13  14  11  11  11  19
 20  17  14  15  15  20  15  17  15  18
 23  20  14  17  19  17  19  20  14  17
 21  13   8  14  20   9  15   8   8  13
 21  15  12  14  24  14  18  14   9  17
 18  24  22  14  28  11  27  19  12  17
-
```

```
BASI
>LINK '5;LSTAT;FRIED'
       MODIFIED  FEB. 74

       DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?*FRIED

       JOB TITLE?TEST BAG SAMPLING SITES

       NUMBER OF:
          OBSERVATIONS / REPLICATION?6
          COLUMNS (TREATMENTS)?10
          REPLICATIONS           ?6
```

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

DATA:

           1      2      3      4      5      6      7      8      9     10
--------------------------------------------------------------------------

         REPLICATION 1
  1--      5      5      3      4      5      3      6      4      6     10
  2--      4      5      3      5      5      4      5      4     11     11
  3--      6      5      2      7      5      3      5      6     10     10
  4--     12      8      2      4      5      4      8      9      6      8
  5--     13     14      3      4      6      5      9     14     10      9
  6--      7     10      3      3      5      4      8     13      9      8

         REPLICATION 2
  1--      8      5      4      7      4      8      7      5      5      8
  2--     12     12      6      8      6     10      7      9     10     12
  3--     13      8      8      8      8      7      7      7      8      9
  4--      7      8      3      4      6      7     10     10      5     11
  5--     15     13      6      4      7      9     14     14      8     13
  6--     12     10      6      7      8      5      6     14     11      9

         REPLICATION 3
  1--     13      8      3      6      7      8     10      8      5     11
  2--     20     11      4     13     10     10     16     14     10     13
  3--     13     13      6      8     10      8     10     10      9      9
  4--     13      6      5      8      6      9     10     10      8     14
  5--     17      8      9     10      8     14     12     14     11     21
  6--     14     10      8     10     11      8      7      8     10     13

         REPLICATION 4
  1--      8      8      3      4      5      5     12     10      7      9
  2--     17     14      3      5      8      6     13     14      8     11
  3--      8     11      4      5      5      6      9     18      8      8
  4--     20     12      8     15     14     14     15     13      8     13
  5--     21     16     16     13     20     16     18     23     11     19
  6--     14     13     10     11     10     14     14     12     14     13

         REPLICATION 5
  1--     21     18      9     14     21     12     14     16      7     13
  2--     33     25     14     22     25     14     16     22     11     16
  3--     16     16     10     13     22     15     19     16     11     15
  4--     30     13     15     22     24     17     24     19     16     24
  5--     34     10     24     29     27     30     30     21     21     20
  6--     27      8     21     25     22     22     21     22     17     18

         REPLICATION 6
  1--     13     12     10     11     13     14     11     11     11     19
  2--     20     17     14     15     15     20     15     17     15     18
  3--     23     20     14     17     19     17     19     20     14     17
  4--     21     13      8     14     22      9     15      8      8     13
  5--     21     15     12     14     24     14     18     14      9     17
  6--     18     24     22     14     28     11     27     19     12     17
```

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

INPUT NUMBER OF COLUMNS IN ANALYSIS?10


RANKS

            1      2      3      4      5      6      7      8      9     10
--------------------------------------------------------------------------------

            REPLICATION 1
 1--   14.0   14.0    2.5    7.0   14.0    2.5   20.0    7.0   20.0   38.5
 2--    4.5    9.0    1.5    9.0    9.0    4.5    9.0    4.5   27.0   27.0
 3--   10.5    6.0    1.0   14.5    6.0    2.0    6.0   10.5   33.5   33.5
 4--   34.5   20.0    1.0    4.0    7.0    4.0   20.0   27.0   10.5   20.0
 5--   24.5   31.0    1.0    2.5    5.5    4.0   13.0   31.0   17.0   13.0
 6--    9.0   23.0    1.5    1.5    4.5    3.0   14.0   35.5   18.5   14.0

            REPLICATION 2
 1--   30.5   14.0    7.0   24.0    7.0   30.5   24.0   14.0   14.0   30.5
 2--   31.0   31.0   13.0   17.0   13.0   22.0   15.0   19.0   22.0   31.0
 3--   40.5   21.5   21.5   21.5   21.5   14.5   14.5   14.5   21.5   28.5
 4--   13.5   20.0    2.0    4.0   10.5   13.5   30.5   30.5    7.0   33.0
 5--   36.5   24.5    5.5    2.5    7.0   13.0   31.0   31.0    9.0   24.5
 6--   32.0   23.0    6.5    9.0   14.0    4.5    6.5   41.0   28.5   18.5

            REPLICATION 3
 1--   50.5   30.5    2.5   20.0   24.0   30.5   38.5   30.5   14.0   43.0
 2--   54.0   27.0    4.5   34.0   22.0   22.0   47.0   38.5   22.0   34.0
 3--   40.5   40.5   10.5   21.5   33.5   21.5   33.5   33.5   28.5   28.5
 4--   38.0   10.5    7.0   20.0   10.5   27.0   30.5   30.5   20.0   42.5
 5--   41.5    9.0   13.0   17.0    9.0   31.0   21.5   31.0   19.5   50.0
 6--   41.0   23.0   14.0   23.0   28.5   14.0    9.0   14.0   23.0   35.5

            REPLICATION 4
 1--   30.5   30.5    2.5    7.0   14.0   14.0   47.0   38.5   24.0   35.5
 2--   50.0   38.5    1.5    9.0   17.0   13.0   34.0   38.5   17.0   27.0
 3--   21.5   37.5    3.0    6.0    6.0   10.5   28.5   53.0   21.5   21.5
 4--   53.0   34.5   20.0   46.5   42.5   42.5   46.5   38.0   20.0   51.0
 5--   50.0   39.0   39.0   24.5   46.5   39.0   43.5   53.0   19.5   45.0
 6--   41.0   35.5   23.0   28.5   23.0   41.0   41.0   32.0   41.0   35.5

            REPLICATION 5
 1--   59.5   57.0   35.5   54.0   59.5   47.0   54.0   56.0   24.0   50.5
 2--   60.0   58.5   38.5   56.5   58.5   38.5   47.0   56.5   27.0   47.0
 3--   48.0   48.0   33.5   40.5   59.0   45.5   55.0   48.0   37.5   45.5
 4--   60.0   38.0   46.5   55.5   58.0   50.0   58.0   52.0   49.0   58.0
 5--   60.0   17.0   54.5   57.0   56.0   58.5   58.5   50.0   50.0   46.5
 6--   58.5   14.0   50.5   57.0   53.5   53.5   50.5   53.5   45.5   47.5

            REPLICATION 6
 1--   50.5   47.0   38.5   43.0   50.5   54.0   43.0   43.0   43.0   58.0
 2--   54.0   50.0   38.5   43.5   43.5   54.0   43.5   50.0   43.5   52.0
 3--   60.0   57.5   43.5   51.0   55.0   51.0   55.0   57.5   43.5   51.0
 4--   54.0   38.0   20.0   42.5   55.5   27.0   46.5   20.0   20.0   38.0
 5--   50.0   36.5   21.5   31.0   54.5   31.0   43.5   31.0   13.0   41.5
```

```
6--  47.5  56.0  53.5  41.0  60.0  28.5  58.5  49.0  32.0  45.5
```

BENNETT

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

COLUMN RANK TOTALS:

    COLUMN  1 = 1454.5
    COLUMN  2 = 1110.5
    COLUMN  3 =  679.0
    COLUMN  4 =  946.0
    COLUMN  5 = 1059.0
    COLUMN  6 =  962.5
    COLUMN  7 = 1237.0
    COLUMN  8 = 1263.0
    COLUMN  9 =  926.5
    COLUMN 10 = 1342.0

SIGNIFICANCE LEVEL?.05


CHI-SQUARE =    43.833     DF =    9     PROBABILITY = .000[illegible]

        REJECT HYPOTHESIS THAT THE 10 DISTRIBUTIONS ARE
FROM THE SAME POPULATION.


ANOTHER ANALYSIS USING THE SAME DATA?YES
```

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

INPUT NUMBER OF COLUMNS IN ANALYSIS?3
INPUT 3 COLUMN NUMBERS IN ASCENDING ORDER
?4,6,9


RANKS

           4       6       9
----------------------------


           REPLICATION 1
 1--     2.5     1.0     7.5
 2--     2.5     1.0    11.5
 3--     4.5     1.0    12.0
 4--     2.0     2.0     5.0
 5--     1.5     3.0     7.5
 6--     1.0     2.0     6.0

           REPLICATION 2
 1--    10.0    12.5     5.0
 2--     5.5     8.5     8.5
 3--     8.0     4.5     8.0
 4--     2.0     6.0     4.0
 5--     1.5     5.5     4.0
 6--     4.0     3.0    10.0

           REPLICATION 3
 1--     7.5    12.5     5.0
 2--    13.0     8.5     8.5
 3--     8.0     8.0    11.0
 4--     8.5    11.5     8.5
 5--     7.5    13.0     9.5
 6--     7.5     5.0     7.5

           REPLICATION 4
 1--     2.5     5.0    10.0
 2--     2.5     4.0     5.5
 3--     2.0     3.0     8.0
 4--    15.0    13.5     8.5
 5--    11.0    15.0     9.5
 6--    10.0    14.0    14.0

           REPLICATION 5
 1--    17.5    16.0    10.0
 2--    18.0    14.0    11.5
 3--    14.0    16.0    13.0
 4--    18.0    17.0    16.0
 5--    17.0    18.0    16.0
 6--    18.0    17.0    16.0

           REPLICATION 6
 1--    14.5    17.5    14.5
 2--    15.5    17.0    15.5
 3--    17.5    17.5    15.0
```

```
4--  13.5  11.5   8.5
5--  13.0  13.0   5.5
6--  14.0  10.0  12.0
```

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST-BAG SAMPLING SITES --

COLUMN RANK TOTALS:

  COLUMN  4 =  330.5
  COLUMN  6 =  347.5
  COLUMN  9 =  348.0

SIGNIFICANCE LEVEL?.05


CHI-SQUARE =      .193     DF =   2     PROBABILITY = .90780

 DO NOT REJECT HYPOTHESIS THAT THE  3 DISTRIBUTIONS ARE
FROM THE SAME POPULATION.


ANOTHER ANALYSIS USING THE SAME DATA?YES
```

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

INPUT NUMBER OF COLUMNS IN ANALYSIS?4
INPUT 4 COLUMN NUMBERS IN ASCENDING ORDER
?2,7,8,10


RANKS

             2         7         8        10
--------------------------------------------


                 REPLICATION 1
 1--       3.0       5.0       1.0      13.0
 2--       2.5       2.5       1.0       7.0
 3--       1.5       1.5       3.0      12.0
 4-        4.0       4.0       7.0       4.0
 5-       10.5       2.5      10.5       2.5
 6--       9.0       4.5      13.5       4.5

                 REPLICATION 2
 1--       3.0       6.0       3.0       8.5
 2--       9.5       4.0       5.0       9.5
 3-        6.5       4.5       4.5       9.0
 4--       4.0       9.5       9.5      12.0
 5--       6.5      10.5      10.5       6.5
 6--       9.0       1.0      16.5       7.0

                 REPLICATION 3
 1--       8.5      13.0       8.5      16.0
 2--       7.0      18.0      14.0      11.5
 3--      15.0      12.0      12.0       9.0
 4--       1.0       9.5       9.5      18.0
 5--       1.0       5.0      10.5      21.5
 6--       9.0       2.0       4.5      13.5

                 REPLICATION 4
 1--       8.5      18.5      13.0      11.0
 2-       14.0  [illegible]   14.0       7.0
 3--      14.0       9.0      20.0       6.5
 4--      13.0      19.5      15.5      21.0
 5--      15.0      17.5      23.0      19.0
 6-       13.5      16.5      11.0      13.5

                 REPLICATION 5
 1-       23.0      21.0  [illegible]   22.0
 2-       24.0      18.0      23.0      18.0
 3--      17.5      21.5      17.5      16.0
 4-       .15.      23.5  [illegible]   23.5
 5--       4.0      24.0      21.5      20.0
 6--       4.5      21.0      22.0      19.0

                 REPLICATION 6
 1--      18.5      16.0      16.0      24.0
 2--      20.5      16.0      20.5      22.0
 3-       24.5      21.5      23.5      19.0
 4--      15.5      19.5       4.0      15.5
```

5--	14.0	17.5	10.5	16.0
6--	23.0	24.0	20.0	18.0

```
FRIEDMAN 2-WAY ANOVA BY RANKS

TEST BAG SAMPLING SITES

COLUMN RANK TOTALS:

   COLUMN  2 =  392.0
   COLUMN  7 =  451.0
   COLUMN  8 =  463.0
   COLUMN 10 =  494.0

SIGNIFICANCE LEVEL?.05


CHI-SQUARE =    3.039     DF =   3     PROBABILITY = .38567

 DO NOT REJECT HYPOTHESIS THAT THE  4 DISTRIBUTIONS ARE
FROM THE SAME POPULATION.


ANOTHER ANALYSIS USING THE SAME DATA?NO
>
```

CHI-SQUARED TEST

The computer program which runs the Chi-squared test is called "5;LSTAT;CONTIN". This program tests the sampling distribution (of air or meteorological data) of a short field sampling period and compares it to the historical distribution of the same variable. This could be used to compare the distribution of one year's data, for example, with the distribution of the previous 10 years data at the same location to see if the year under consideration could be considered a "typical" year.

This program can handle a data matrix of up to 30 columns by 30 rows. The information that is needed for this test can be obtained from yearly and 10-year summaries respectively of air or meteorological data. The test should be run for the frequency of occurrence of wind directions and repeated at least for wind speed classes and stability before observance of a "typical" year can be determined. See previous discussion for complete details.

The summary that should be used to determine whether the year was "typical" is the one for all stability classes combined. This should be for the whole season and can be made up by adding the total columns from the monthly summary in the "WIND2", "STAR2" or "WINDROS" computer printouts if necessary. Page 137 of the Meteorology manual will be used to show where the data is lifted from the computer summary. Refer to the printout for the Hayward Airport. This is an example of part of the data required for the long term sample.

From that summary page, the 1966-1970 data was obtained. The 1972 data was obtained from a similar summary sheet for the year 1972 alone.

Direction	1966 - 1970 Number of Occurrences	1972 Number of Occurrences
N	30	7
NNE	28	4
NE	61	12
ENE	152	30
E	86	16
ESE	65	15
SE	78	18
SSE	49	10
S	23	4
SSW	22	3
SW	32	6
WSW	56	10
W	97	22
WNW	52	10
NW	53	11
NNW	32	6
*	*	*
TOTAL	916	184

Speed	1966 - 1970 Number of Occurrences	1972 Number of Occurrences
0-3	76	18
4-7	552	108
8-12	218	46
13-18	53	10
19-24	9	2
24	7	0
TOTAL	915	184

*Note: Do not include calms if they have been distributed.

In an actual test, the long term data would be combined with data for the rest of the months and for each time period.

```
-COPY CHI%DIRE TO TEL TEXT



HAYWARD AIRPORT 1966-1970 DATA VS 1972 DATA DIRECTION,16,2
  30   7
  28   4
  61  12
 152  30
  86  16
  65  15
  78  18
  49  10
  23   4
  22   3
  32   6
  56  10
  97  22
  52  10
  53  11
  32   6
-
```

```
COPY CHI%SPED TO TEL TEXT



HAYWARD AIRPORT 1966-1970 DATA VS 1972 DATA SPEED,6,2
  76  18
 552 108
 218  46
  53  10
   9   2
   7   0
-
```

```
BASI
>LINK '5;LSTAT;CONTIN'
DATA FILENAME OR 'EXP' FOR INPUT FORMAT?CHI%DIRE
```

```
  CHI-SQUARE CONTINGENCY TABLE...HAYWARD AIRPORT 1966-1970 DATA VS 1972
DATA DIRECTION

  CELL          ACTUAL         EXPECTED

  1,1           30             30.8109
  1,2           7              6.18909
  2,1           28             26.6473
  2,2           4              5.35273
  3,1           61             60.7891
  3,2           12             12.2109
  4,1           152            151.556
  4,2           30             30.4436
  5,1           86             84.9382
  5,2           16             17.0618
  6,1           65             66.6182
  6,2           15             13.3818
  7,1           78             79.9418
  7,2           18             16.0582
  8,1           49             49.1309
  8,2           10             9.86909
  9,1           23             22.4836
  9,2           4              4.51636
  10,1          22             20.8182
  10,2          3              4.18182
  11,1          32             31.6436
  11,2          6              6.35636
  12,1          56             54.96
  12,2          10             11.04
  13,1          97             99.0945
  13,2          22             19.9055
  14,1          52             51.6291
  14,2          10             10.3709
  15,1          53             53.2945
  15,2          11             10.7055
  16,1          32             31.6436
  16,2          6              6.35636

  CHI-SQUARE (15 D.F.) = 2.0766


>
```

```
RUN
DATA FILENAME OR 'EXP' FOR INPUT FORMAT?CHI%SPED
```

```
 CHI-SQUARE CONTINGENCY TABLE...HAYWARD AIRPORT 1966-1970 DATA VS 1972
DATA SPEED

 CELL       ACTUAL       EXPECTED

 1,1        76           78.2621
 1,2        18           15.7379
 2,1        552          549.5
 2,2        108          110.5
 3,1        218          219.8
 3,2        46           44.2002
 4,1        53           52.4522
 4,2        10           10.5478
 5,1        9            9.15833
 5,2        2            1.84167
 6,1        7            5.82803
 6,2        0            1.17197

 CHI-SQUARE (5 D.F.) = 2.00466


>
```

The data is entered into the computer in the same manner as described for the CATANOVA test. The job title, the number of rows, and the number of columns are entered on the first line of the data file. The rest of the datafile is the same as for the CATANOVA test.

The basic mode is entered, and the program is started by typing LINK '5;LSTAT;CONTIN'. The computer asks for either the datafile name (CHI%DIRE in this example) or EXP if program information is required.

The computer prints out the title of the statistical test followed by the job title. Next the computer prints out the cell designation, the actual frequency of occurance as exists in the data file, and the expected frequency of occurance according to statistical laws. The calculated chi-square is printed along with the associated degrees of freedom. In this example, the calculated chi-square is 2.0766 with 15 degrees of freedom. The statistical value of chi-square with a 5 percent level of significance is 25.00. Since the calculated chi-square is less than the statistical chi-square $(2.0766 < 25.00)$ accept the hypothesis that the short term sample is the same as the long term sample.

Question: What are the following values for the wind speed classes?

$\alpha = .05$

Calculate chi-square =

Statistical chi-square =

Degrees of freedom =

Conclusion:

REGRESSION TECHNIQUES

REGRESSION ANALYSIS STEP BY STEP PROCEDURE

1. Collect data.

2. Determine physical relationships.

3. Plot data - scatter diagram.

4. Use least squares to develop bestline.

5. Use F-test to check significance of entire equation.

6. Use t-test to check the significance of the regression coefficients.

7. Check the highest R-value of equations.

8. Check the smallest $S_{\bar{y}x}$ (standard error).

9. Select best equation from Steps 5, 6, 7 and 8 above.

10. Consider transformation of data if necessary.

*11. Check assumptions - normality and constant variance by plotting the residuals.

*Consider Step 11 only if all Steps above are passed.

LINEAR REGRESSION

Regression techniques can be used to perform several functions. One such application that can be used is the comparison of bag sampling techniques with continuous monitoring at fixed stations such as APCD's. A nonparametric test which can be used if there is not a very large range in the data and there are not a large amount of data pairs is described in the section titled "Wilcoxon Matched Sign Test". If there are a relatively large number of data pairs, say 15 or more, and the range is fairly large say 4 or greater, then an attempt should be made to use regression techniques. In addition to telling whether the two sets of observations are identical, a regression equation will be obtained, enabling the data to be transformed from one sampling technique to comparable values for the other sampling technique. Since the purpose of the test under consideration is to compare one set of data with one other set of data, a two-dimensional linear test can be employed. The computer name of this linear test is "5;LSTAT;LINREG". The following is a sample of the data that is required for this test:

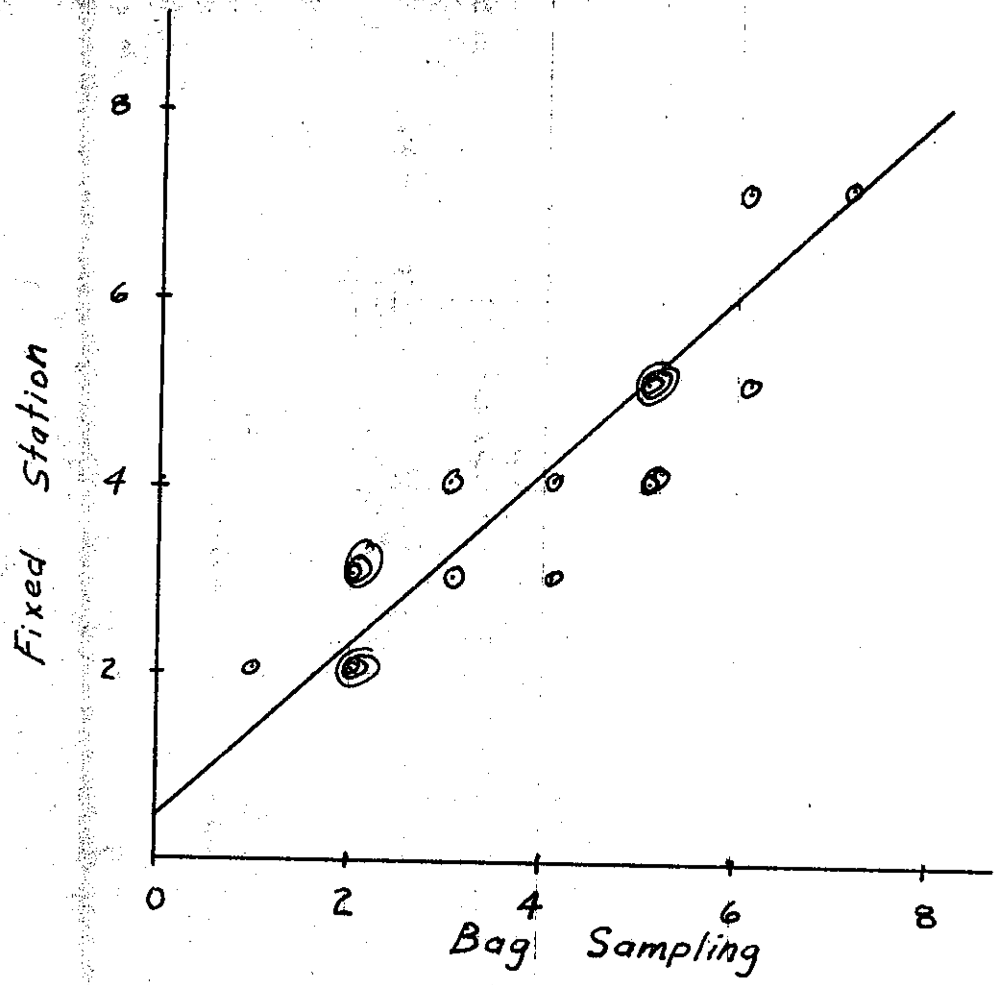
TIME	BAG SAMPLING	FIXED STATION
0700	3	3
0800	4	5
0900	4	5
1000	5	5
1100	5	6
1200	7	7
1300	7	6
1400	4	4
1500	3	4
0700	2	2
0800	2	2
0900	2	1
1000	3	2
1100	3	4
1200	5	5
1300	5	5
0800	2	2
0900	2	3
1000	2	3

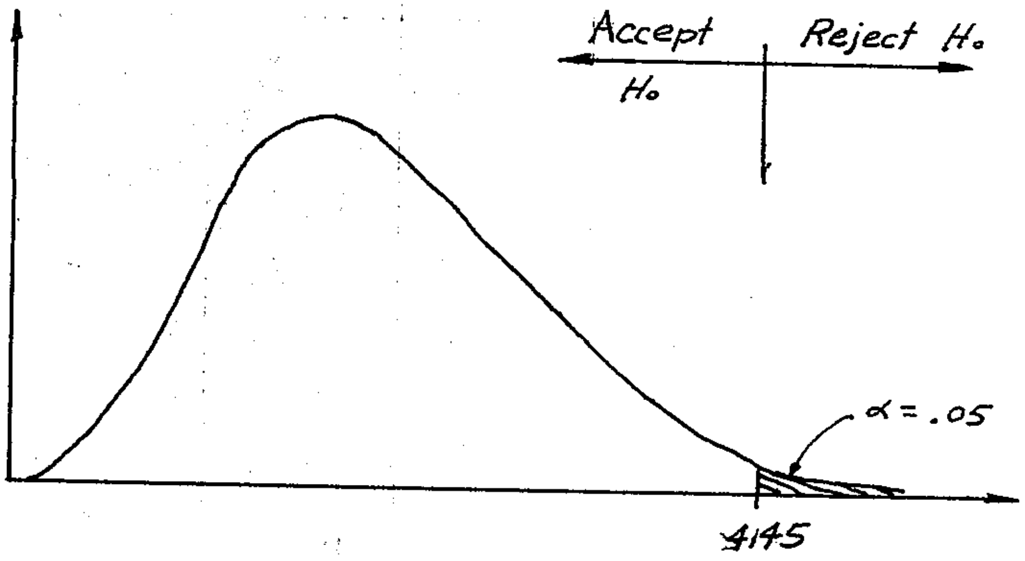
```
-COPY REGRESS TO TEL TEXT


COMPARING APCD SAMPLING WITH BAG SAMPLING,19
   3   3
   4   5
   4   5
   5   5
   5   6
   7   7
   7   6
   4   4
   3   4
   2   2
   2   2
   2   1
   3   2
   3   4
   5   5
   5   5
   2   2
   2   3
   2   3
-
```

```
-BASI
>LINK '5;LSTAT;LINREG'

            ***** MODIFIED 12/6/72 *****

  INPUT DATA FILENAME OR "EXP" FOR PROGRAM EXPLANATION?REGRESS

  DATA TRANSFORMATION CODES:
     0 - NO TRANSFORMATION OF DATA
     1 - V = LOG10(V)
     2 - V = 1/V
  ENTER TRANSFORMATION CODES FOR X,Y?0,0

  FILE FORMAT:
     1 = ALL X VALUES, THEN ALL Y VALUES
     2 = XY PAIRS
     WHICH?2

  WHEN ASKED "WHAT NEXT?" ENTER:
     0   FOR NO LISTING OF X, Y-ACTUAL AND Y-CALCULATED
     1   FOR LISTING WITH DIFFERENCES AND % DIFFERENCES
     2   FOR LISTING WITH 95% CONFIDENCE LIMITS OF YBAR
     3   FOR LISTING WITH 95% PREDICTION LIMITS FOR Y
```

```
LEAST SQUARES LINEAR REGRESSION ANALYSIS

COMPARING APCD SAMPLING WITH BAG SAMPLING

REGRESSION EQUATION: Y=.496717+.922319*X

NUMBER OF OBSERVATIONS        = 19
STD. ERROR OF Y ON X          = .722236
INDEX OF DETERMINATION        = .821898
COEFF. OF CORRELATION         = .906586

SOURCE                 SS          DF              MS          F

REGRESSION             40.9219     1               40.9219     78.4509
RESIDUAL               8.86761     17              .521624

TOTAL                  49.7895     18

VAR.      MEAN         VARIANCE    STD. DEV.

   X      3.68421      2.67251     1.63478
   Y      3.89474      2.76608     1.66315

          STD. ERROR   95% CONFIDENCE LIMITS

YINTCPT   .417894      -.384973    1.37841
SLOPE     .104132      .702618     1.14202

WHAT NEXT?2
```

```
LEAST SQUARES LINEAR REGRESSION ANALYSIS

COMPARING APCD SAMPLING WITH BAG SAMPLING

REGRESSION EQUATION: Y=.496717+.922319*X

          X-ACTUAL     Y-ACTUAL     Y-CALC       95% CONFIDENCE LIMITS

          3            3            3.26368      2.88314      3.64421
          4            5            4.186        3.82959      4.5424
          4            5            4.186        3.82959      4.5424
          5            5            5.10831      4.65469      5.56194
          5            6            5.10831      4.65469      5.56194
          7            7            6.95295      6.14493      7.76097
          7            6            6.95295      6.14493      7.76097
          4            4            4.186        3.82959      4.5424
          3            4            3.26368      2.88314      3.64421
          2            2            2.34136      1.83231      2.8504
          2            2            2.34136      1.83231      2.8504
          2            1            2.34136      1.83231      2.8504
          3            2            3.26368      2.88314      3.64421
          3            4            3.26368      2.88314      3.64421
          5            5            5.10831      4.65469      5.56194
          5            5            5.10831      4.65469      5.56194
          2            2            2.34136      1.83231      2.8504
          2            3            2.34136      1.83231      2.8504
          2            3            2.34136      1.83231      2.8504

WHAT NEXT?0
```


Fixed Station
Bag Sampling
0
2
4
6
8
2
4
6
8


Accept H₀
Reject H₀
α = .05
4145

The data file that is required to run this statistical analysis by computer is shown next. The first line of the data file consists of a descriptive job title selected by the user, followed by the number of data pairs to be analyzed. The program is started by entering the "basic" mode and typing LINK '5;LSTAT;LINREG'.

The program asks for the data file name or EXP for program explaination. In this example, the data file name, REGRESS is entered. The data is not transformed in this example, so 0,0 (zero, zero) is entered for the data transformation codes. It is good practice to run this program with no transformation and then to repeat the test later with 1,1 as the transformation codes. This results in a log-log regression analysis. The program next asks for the file format. Since the file was constructed as "xy" pairs, a 2 is entered.

The program then indicates which code (0, 1, 2, or 3) to use at the end of the run. It then advances the page and prints the title of the statistical test followed by the job title selected by the user. Next the regression equation which yields the best fit is printed. In this example the regression equation is Y=.496717+.922319*X. The numbers of observation (N), the standard error of the estimate ($S_{\overline{y}}$), the index of determination (r^2) and the coefficient of correlation (r) are all printed. For this example these values are the following:

N = 19

$S_{\overline{y}}$ = .722236 (say .72

r^2 = .821898 (say .82

r = .906586 (say .91)

Next the computer prints out the data required to assess the variability of the data. These lead to the calculated "F-ratio" and the degrees of freedom which are the following in this example:

F = 78.4509 (say 78.45)

Degrees of freedom = 1,17

The calculated "F-ratio" is compared to the statistical "F-ratio" in tables. Use a level of significance of 5 percent and the degrees of freedom indicated. With 1,17 degrees of freedom at a level of significance of 5 percent, the statistical "F-ratio" is 4.45. Since the calculated "F-ratio" is greater than the statistical "F-ratio", the relationship between the two sets of data are significant. This means that we can predict a value associated with the "y" set of data solely from one measurement from the "x" location. (The "y" data are the dependent data.)

The computer next prints the mean, variance and the standard deviation of both the "x" and "y" sets of data. These should be checked to see if they appear reasonable. Almost all of the data should fall within a range of the mean minus two standard deviations and the mean plus two standard deviations.

Next the computer tests the regression equation to see if the data sets are statistically alike. The "y" intercept confidence limits, -.384973 and 1.37841 should "span" zero. (One should be negative and one positive.) The confidence interval for the slope should "span" 1.0 (one value should be less than 1.0 and one value should be greater than 1.0). If both of the above two conditions are met, there is, statistically speaking, no difference between the two sets of data.

The computer then asks WHAT NEXT?. This is the place to input the code listed before the computer started the statistical analysis. It is recommended that the code "2" be used in order to get the confidence limits of the regression line. This was done in this example. The computer moved to a new page and repeated the job title and the regression equation. It then proceded to print out the actual x value, the actual y value, the y value from the regression equation, and the lower and upper confidence limits respectively for each observation in the data file. At the end of this procedure, the computer once again asked WHAT NEXT?. As no further information was required, a "O" was entered.

This test should be performed three times. When the initial data is collected and before any further field sampling is conducted, run this test to verify calibration, instruments, etc. The comparative data should be obtained twice more, once about half way through the period of sampling and again when all sampling is complete. All three sets of data should be combined at the end of the sampling period to obtain an overall check. If discrepancies are noted at this time, the data should be run separately to see when the problem occurred.

STEPWISE MULTIPLE LINEAR REGRESSION

This statistical technique can be of value when continuous sampling is not possible at a site. For CO, continuous multiple bag sampling is the best approach. However, when a mobile van is required at more than one location, this technique can be of help for filling in the data set. A regression equation is used to estimate the field data from other data which is continuously measured.

The derived relationship can be used to estimate the field concentrations for time periods during which field measurements were not made, as long as similar meteorological and traffic conditions exist.

An example of the data required for this test follows.

		Van	APCD	Sky	Inversion	Wind
Date	Time	Ozone	Ozone	Code	Height	Speed
6-11-73	0600	.03	.04	1	5	2
"	0700	.03	.04	1	7	2
"	0800	.03	.03	2	10	3
"	0900	.05	.04	2	12	5
"	1000	.05	.04	2	12	5
"	1100	.05	.04	2	14	6
"	1200	.05	.05	2	16	7
"	1300	.07	.06	2	20	7
"	1400	.07	.06	2	25	7
"	1500	.06	.06	2	25	6
"	1600	.05	.06	2	27	5
"	1700	.03	.04	3	28	4
"	1800	.03	.04	2	30	4
"	1900	.02	.03	2	32	2
6-12-73	0700	-	.01	0	6	2
"	0700	-	.01	0	6	2
"	0800	-	.02	1	8	3
"	0900	-	.02	1	8	3
"	1000	-	.02	1	9	3
"	1100	-	.03	2	10	4
"	1200	-	.02	2	12	2
"	1300	-	.02	2	12	2
"	1400	-	.02	2	16	2
"	1500	-	.01	3	16	2
"	1600	-	.01	2	16	2
"	1700	-	.01	2	18	2
"	1800	-	.01	2	20	1
"	1900	-	.01	1	20	1
6-13-73	0600	.01	.02	0	15	2
"	0700	.02	.02	0	10	2
"	0800	.02	.01	1	15	3
"	0900	.02	.02	1	15	3
"	1000	.03	.03	1	16	3
"	1100	.03	.02	1	16	3
"	1200	.04	.02	1	17	5
"	1300	.05	.04	2	20	5
"	1400	.07	.06	3	23	7
"	1500	.06	.07	3	25	7
"	1600	.06	.04	2	25	6
"	1700	.04	.03	3	30	4
"	1800	.03	.02	3	35	4
"	1900	.03	.02	3	35	3
6-14-73	0600	-	.01	0	15	2
"	0700	-	.01	0	10	2
"	0800	-	.01	0	10	2
"	0900	-	.02	2	12	3
"	1000	-	.02	2	16	4
"	1100	-	.02	3	18	5
"	1200	-	.02	4	20	5
"	1300	-	.04	4	20	6
"	1400	-	.02	3	20	5
"	1500	-	.02	2	24	4
"	1600	-	.02	2	24	4
"	1700	-	.02	2	22	4
"	1800	-	.01	1	22	3
"	1900	-	.01	1	22	2
6-15-73	0600	.02	.02	4	7	2
"	0700	.02	.02	4	9	2
"	0800	.03	.02	4	10	3
"	0900	.03	.02	4	15	4
"	1000	.04	.03	5	20	6
"	1100	.05	.03	5	25	6
"	1200	.04	.03	4	25	6
"	1300	.04	.03	4	25	6
"	1400	.04	.03	4	27	5
"	1500	.03	.03	4	30	5
"	1600	.03	.02	2	25	5
"	1700	.02	.02	2	24	3
"	1800	.01	.01	2	20	3
"	1900	.01	.01	1	17	2

The following is a listing of the data used in the derivation of the relationship between the van ozone measurements and the other data as described above.

```
-COPY STEP TO TEL TEXT


06 .03 .04 1  5 2
07 .03 .04 1  7 2
08 .03 .03 2 10 3
09 .05 .04 2 12 5
10 .05 .04 2 12 5
11 .05 .04 2 14 6
12 .05 .05 2 16 7
13 .07 .06 2 20 7
14 .07 .06 2 25 7
15 .06 .06 2 25 6
16 .05 .06 2 27 5
17 .03 .04 3 28 4
18 .03 .04 2 30 4
19 .02 .03 2 32 2
06 .01 .02 0 10 2
07 .02 .02 0 10 2
08 .02 .01 1 15 3
09 .02 .02 1 15 3
10 .03 .03 1 16 3
11 .03 .02 1 16 3
12 .04 .02 1 17 5
13 .05 .04 2 20 5
14 .07 .06 3 23 7
15 .06 .07 3 25 7
16 .06 .04 2 25 6
17 .04 .03 3 30 4
18 .03 .02 3 35 4
19 .03 .02 3 35 3
06 .02 .02 4  7 2
07 .02 .02 4  9 2
08 .03 .02 4 10 3
09 .03 .02 4 15 4
10 .04 .03 5 20 6
11 .05 .03 5 25 6
12 .04 .03 4 25 6
13 .04 .03 4 25 6
14 .04 .03 4 27 5
15 .03 .03 4 30 5
16 .03 .02 2 25 5
17 .02 .02 2 24 3
18 .01 .01 2 20 3
19 .01 .01 1 17 2
-
```

To run the Stepwise Multiple Linear Regression test, enter the Basic mode (>) and type LINK'5;LSTAT;STPREG'. When asked for data filename or EXP for program details, type EXP the first time this program is used. This will initiate a printout of about 5 pages of explanation. This is good reference material which can be used if difficulties are encountered.

When ready to run the program, input the name of the datafile which contains the data (STEP in this example). After "JOB TITLE?" is printed by the computer, enter any descriptive job title. This number of variables is the same as the number of columns in the datafile. Always type in NO when asked if a forced zero intercept is required.

The computer next asks the user to label the variables with a name which is no longer than 10 characters in length. The variable noted as X(1) is the variable in the first column of the datafile, etc. In this example, the time of day and height to the inversion base were divided by 100 so that the computer would have room to print the coefficient of these variables (they are low numbers).

The computer will next printout the above input data, plus the mean, variance and standard deviation for each variable. This is similar to the information in '5;LSTAT;LINREG'. Next the correlation matrix is printed. This given an idea of the per-cent of correlation that each variable has with each of the other variables.

Control data is asked for next.

The F - level for inclusion and deletion and the Tolerance Level are requested. For the first run of a set of data, enter "O" (zero) for all three.

```
BASI
>LINK '5;LSTAT;STPREG'

          ***** MODIFIED 2/27/73 *****

   DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?STEP

      A WORK FILE '$REG' HAS BEEN CREATED FOR USE BY THIS
   PROGRAM.  DELETE THIS FILE WHEN COMPLETED WITH THIS RUN.

   JOB TITLE?COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA
   NUMBER OF VARIABLES?6
   FORCE ZERO REGRESSION INTERCEPT (YES OR NO)?NO

   LABELS<=10 CHAR. (ENTER NUMERIC "0" FOR BLANK):
      X(1)  =  ?TIME/100
      X(2)  =  ?VAN OZ.
      X(3)  =  ?APCD OZ.
      X(4)  =  ?SKY CODE
      X(5)  =  ?INV/100
      X(6)  =  ?WIND SP.
```

```
 STEPWISE MULTIPLE LINEAR REGRESSION

 COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

 NUMBER OF OBSERVATIONS    42
 NUMBER OF VARIABLES        6
 FORCE ZERO INTERCEPT      NO



 VARIABLE              MEAN         VARIANCE        STD. DEV.

TIME/100  1         12.50000        16.64634          4.07999
 VAN OZ.  2           .03667          .00027           .01633
APCD OZ.  3           .03262          .00023           .01531
SKY CODE  4          2.38095         1.60743          1.26785
 INV/100  5         19.85714        65.10105          8.06852
WIND SP.  6          4.28571         2.89199          1.70058


 CORRELATION MATRIX

 ROW 1
    1.00000

 ROW 2
     .12813    1.00000

 ROW 3
     .13470     .84542    1.00000

 ROW 4
     .10845     .20420     .03530    1.00000

 ROW 5
     .87056     .24250     .18671     .33686    1.00000

 ROW 6
     .25661     .87828     .68246     .35553     .39411    1.00000
```

Control values for the variables asked for next. Also enter a numeral 1 (for the first run of a set of data) for the field measurement that you are trying to predict and a numeral 2 for the rest of the variables. When asked for the number of steps, enter the maximum number that may be required. This control data will give the computer freedom to consider all the data and will give the user an idea of the order of importance of the variables.

The computer next goes through a stepwise process. The computer, at Step 1, enters the independent variable which explains the largest amount of the variation in the dependent variable. The variable which explains the next largest amount of variation is entered next, and so on. This process is repeated until either all the variables are used or until the improvement is insignificant. At this point the computer prints "F-LEVEL OR TOLERANCE INSIGNIFICANT FOR FURTHER COMPUTATION". It then prints a summary of the steps, listing the multiple correlation, standard error, F-ratio and the number of variables that are in the final regression equation.

Look at Problem 1, Step 4. The first five lines are the input data. The variable that was entered in this step was variable number 4, the sky code. The multiple correlation is .94. This means that 94% of the variation in the dependent variable (the van ozone) is explained by the four variables in the regression (time/100,APCD Ozone, sky code, and wind speed).

Next an analysis of variance (ANOVA) table is printed. The line labeled "REGRESSION" is for the variation that is explained by the regression equation and the line labeled "RESIDUAL" is the variation that is in the error term (Made up of measured and

```
CONTROL DATA FOR PROBLEM NO.  1

F-LEVEL FOR INCLUSION (0=.01)?0
F-LEVEL FOR DELETION (0=.005)?0
TOLERANCE LEVEL (0=.001)?0

VARIABLE CONROL VALUES
   0 - DELETE VARIABLE FROM ANALYSIS
   1 - DEPENDENT VARIABLE
   2 - FREE VARIABLE - MAY BE USED IN ANALYSIS
   3 TO 9 - FORCED VARIABLE - LOW TO HIGH LEVEL

CONTROL VALUE FOR:
TIME/100  1 = ?2
 VAN OZ.  2 = ?1
APCD OZ.  3 = ?2
SKY CODE  4 = ?2
 INV/100  5 = ?2
WIND SP.  6 = ?2

THIS PROBLEM MAY REQUIRE UP TO 10 STEPS TO SOLVE.

ENTER THE MAXIMUM NUMBER OF STEPS DESIRED FOR SOLUTION?10
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER            1                    F TO ENTER          .01000
STEP NUMBER               1                    F TO REMOVE         .00500
DEPENDENT VARIABLE        2 (VAN OZ.)          TOLERANCE LEVEL     .00100

VARIABLE ENTERED          6 (WIND SP.)
MULT. CORR. COEFF.         .87828
STD. ERROR EST.            .00791


                              ANOVA TABLE

                  DF        SUM OF SQ.         MEAN SQ.          F-RATIO

REGRESSION         1          .00843            .00843         134.96144
RESIDUAL          40          .00250            .00006
TOTAL             41          .01093



                         VARIABLES IN REGRESSION

  VARIABLE     COEFFICIENT          STD.    COMPUTED     BETA      F-OUT   TYP
                                   ERROR   T- VALUE     COEFF.

  INTERCEPT       .00052
WIND SP.   6      .00843          .00073     11.617       .878    1.3E+02(2)



                      VARIABLES NOT IN REGRESSION

  VARIABLE       PART. CORR.   TOLERANCE     F-IN     TYP

TIME/100   1       -.21044         .934     1.8      (2)
APCD OZ.   3        .70396         .534     3.8E+01(2)
SKY CODE   4       -.24179         .874     2.4      (2)
 INV/100   5       -.23585         .845     2.3      (2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER          1                   F TO ENTER          .01000
STEP NUMBER             2                   F TO REMOVE         .00500
DEPENDENT VARIABLE      2 (VAN OZ.)         TOLERANCE LEVEL     .00100

VARIABLE ENTERED        3 (APCD OZ.)
MULT. CORR. COEFF.       .94057
STD. ERROR EST.          .00569


                           ANOVA TABLE

                DF       SUM OF SQ.        MEAN SQ.         F-RATIO

REGRESSION       2         .00967           .00484        149.58660
RESIDUAL        39         .00126           .00003
TOTAL           41         .01093


                       VARIABLES IN REGRESSION

VARIABLE      COEFFICIENT           STD.    COMPUTED     BETA     F-OUT   TYP
                                   ERROR    T- VALUE    COEFF.

 INTERCEPT        -.00256
APCD OZ.   3       .49114         .07935      6.190      .461    3.8E+01(2)
WIND SP.   6       .00542         .00071      7.581      .564    5.7E+01(2)


                     VARIABLES NOT IN REGRESSION

VARIABLE        PART. CORR.  TOLERANCE     F-IN     TYP

TIME/100   1      -.23997        .931     2.3      (2)
SKY CODE   4      -.04159        .793     6.6E-02(2)
 INV/100   5      -.21230        .832     1.8      (2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER           1                     F TO ENTER           .01000
STEP NUMBER              3                     F TO REMOVE          .00500
DEPENDENT VARIABLE       2 (VAN OZ.)           TOLERNCE LEVEL       .00100

VARIABLE ENTERED         1 (TIME/100)
MULT. CORR. COEFF.        .94410
STD. ERROR EST.           .00559


                               ANOVA TABLE

                  DF       SUM OF SQ.        MEAN SQ.          F-RATIO

REGRESSION         3         .00975           .00325         103.87860
RESIDUAL          38         .00119           .00003
TOTAL             41         .01093


                         VARIABLES IN REGRESSION

   VARIABLE     COEFFICIENT        STD.   COMPUTED     BETA     F-OUT   TYP
                                  ERROR   T- VALUE    COEFF.

  INTERCEPT        .00081
TIME/100   1     -.00034          .00022    -1.524     -.084   2.3      (2)
APCD OZ.   3      .48433          .07816     6.196      .454   3.8E+01(2)
WIND SP.   6      .00567          .00072     7.853      .590   6.2E+01(2)


                      VARIABLES NOT IN REGRESSION

   VARIABLE       PART. CORR.   TOLERANCE     F-IN     TYP

SKY CODE  4         -.04241          .793     6.7E-02(2)
 INV/100  5         -.00873          .206     2.8E-03(2)
```

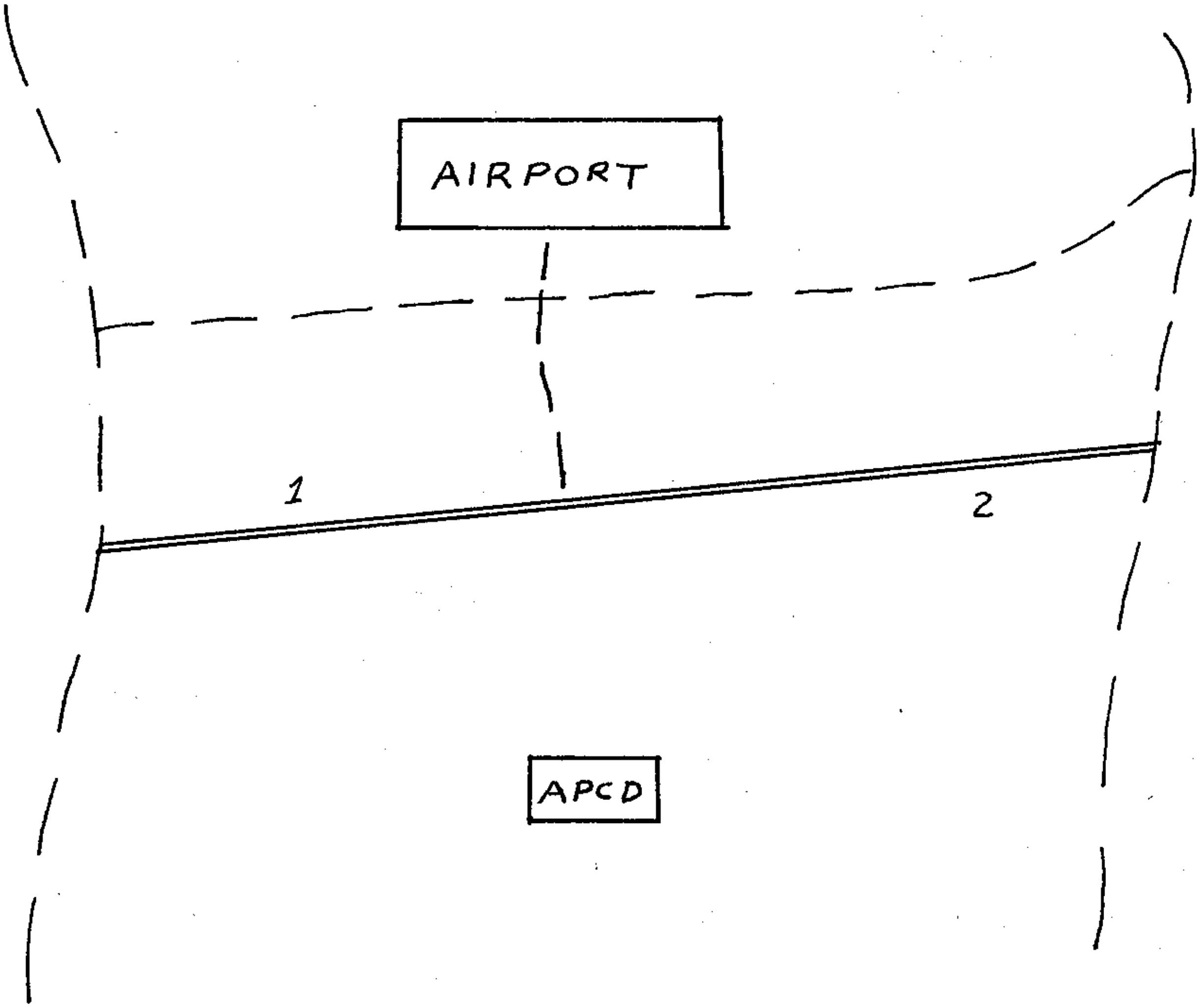
```
  STEPWISE MULTIPLE LINEAR REGRESSION


  COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


  PROBLEM NUMBER              1                   F TO ENTER            .01000
  STEP NUMBER                 4                   F TO REMOVE           .00500
  DEPENDENT VARIABLE          2 (VAN OZ.)         TOLERANCE LEVEL       .00100

  VARIABLE ENTERED            4 (SKY CODE)
  MULT. CORR. COEFF.           .94420
  STD. ERROR EST.              .00566


                                     ANOVA TABLE

                         DF      SUM OF SQ.          MEAN SQ.           F-RATIO

  REGRESSION              4        .00975             .00244           76.01206
  RESIDUAL               37        .00119             .00003
  TOTAL                  41        .01093


                                 VARIABLES IN REGRESSION

   VARIABLE      COEFFICIENT            STD.   COMPUTED     BETA      F-OUT   TYP
                                       ERROR   T- VALUE    COEFF.

    INTERCEPT          .00110
 TIME/100   1        -.00034          .00022     -1.504     -.084    2.3       (2)
 APCD OZ.   3         .47783          .08304      5.754      .448    3.3E+01(2)
 SKY CODE   4        -.00020          .00078     -.258      -.016    6.7E-02(2)
 WIND SP.   6         .00576          .00082      7.064      .600    5.0E+01(2)


                             VARIABLES NOT IN REGRESSION

   VARIABLE         PART. CORR.   TOLERANCE     F-IN     TYP

  INV/100   5           .00921         .173     3.1E-03(2)


   F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION.
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


                                   SUMMARY TABLE

 STEP          VARIABLE          MULT.        STD.              F       NO. IN
NUMBER      NAME    IN OUT       CORR.        ERROR           RATIO       REG.

   1    WIND SP.     6          .87828       .00791       134.96144       1
   2    APCD OZ.     3          .94057       .00569       149.58660       2
   3    TIME/100     1          .94410       .00559       103.87860       3
   4    SKY CODE     4          .94420       .00566        76.01206       4


RESIDUAL ANALYSIS (YES OR NO)?YES
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 1                          DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES    1-TIME/100   ,    3-APCD OZ.   ,    4-SKY CODE
                       6-WIND SP.

 OBS.         ACT.          CALC.         RESIDUAL        NORM.DEV

   1         .03000        .02951         .00049          .08741
   2         .03000        .02917         .00083          .14709
   3         .03000        .02961         .00039          .06922
   4         .05000        .04557         .00443          .78293
   5         .05000        .04523         .00477          .84261
   6         .05000        .05065        -.00065         -.11489
   7         .05000        .06085        -.01085        -1.91633
   8         .07000        .06529         .00471          .83176
   9         .07000        .06495         .00505          .89144
  10         .06000        .05886         .00114          .20215

  11         .05000        .05276        -.00276         -.48715
  12         .03000        .03690        -.00690        -1.21904
  13         .03000        .03677        -.00677        -1.19507
  14         .02000        .02013        -.00013         -.02325
  15         .01000        .02015        -.01015        -1.79278
  16         .02000        .01981         .00019          .03307
  17         .02000        .02025        -.00025         -.04480
  18         .02000        .02469        -.00469         -.82904
  19         .03000        .02913         .0087          .15288
  20         .03000        .02402         .00598         1.05649

  21         .04000        .03520         .00480          .84796
  22         .05000        .04422         .00578         1.02166
  23         .07000        .06475         .00525          .92715
  24         .06000        .06919        -.00919        -1.62326
  25         .06000        .04896         .01104         1.94969
  26         .04000        .03212         .00788         1.39105
  27         .03000        .02701         .00299          .52850
  28         .03000        .02091         .00909         1.60537
  29         .02000        .01934         .00066          .11624
  30         .02000        .01900         .00100          .17592

  31         .03000        .02443         .00557          .98458
  32         .03000        .02985         .00015          .02707
  33         .04000        .04560        -.00560         -.98968
  34         .05000        .04527         .00473          .83617
  35         .04000        .04513        -.00513         -.90602
  36         .04000        .04479        -.00479         -.84634
  37         .04000        .03869         .00131          .23053
  38         .03000        .03836        -.00836        -1.47595
  39         .03000        .03364        -.00364         -.64377
  40         .02000        .02179        -.00179         -.31587
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 1                        DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES    1-TIME/100  ,   3-APCD OZ.  ,   4-SKY CODE
                       6-WIND SP.

  OBS.         ACT.          CALC.        RESIDUAL       NORM.DEV

   41        .01000         .01667        -.00667       -1.17843
   42        .01000         .01078        -.00078        -.13727
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 1                          DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   4-SKY CODE
                      6-WIND SP.

     HISTOGRAM OF RESIDUALS (NORMAL DEVIATES)

 -2 +*
    :*
    :**
    :***
 -1 +**
    :***
    :*
    :**
  0 +*********
    :*****
    :*
    :*****
  1 +*****
    :
    :**
    :
  2 +*


      TEST FOR EXTREME RESIDUAL VALUES   (DIXON CRITERION)

          R22 (MIN) = .089           R22 (MAX) = .156

         R(7)   =     -.011          R(25)   =      .011
        R(15)   =     -.010          R(28)   =      .009
        R(24)   =     -.009          R(26)   =      .008



   DO YOU WISH PLOTS (YES OR NO)?NO
ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)?YES
```


AIRPORT
1
2
APCD

unmeasured variables). The column headed "DF" is the degrees of freedom for the regression and residual respectively. The degrees of freedom associated with the regression is equal to the number of variables in the regression equation. One degree of freedom is "consumed" by each variable. The total degrees of freedom is equal to one less than the total number of observations. The "SUM OF SQUARES" column measures the variability of the regression and residual respectively. When divided by the degrees of freedom, this produces the "MEAN SQUARES". The ratio of the mean squares is the F-ratio.

Next the computer lists a description of the variables that are used in the regression equation. The column labeled "COEFFICIENT" is where the regression equation is listed. For this equation, the regression equation is:

$$\text{Van oz} = 0.0011 - 3.4\text{x}10^{-6}\ (\text{TIME}) + .48\ (\text{APCD OZ})$$
$$- 2.0\text{x}10^{-4}\ (\text{Sky Code}) + 5.76\text{x}10^{-3}\ (\text{Wind Speed})$$

The computed T-value, the standard error and the Beta coefficient are all important in the individual variables which make up the regression equation. Use the computed T-value to test the significance of each variable. If the absolute value of the computed T-value is greater than the tabled value (1.68 with = .05 and 40 degrees of freedom), then the variable has a statistical significance. If the value in the "F-OUT" column is greater than the "F to REMOVE" value selected by the user, the variable will be removed in the next step.

The computer next prints a description of the variables which are available for use in the regression, but have not been used yet. The "F-IN" column dictates when each variable is considered in the regression equation, with the highest F-IN considered first. Note the variable number 5 does not enter since its "F-IN" is less than 0.01.

At this point, review of the Summary Table indicates that the greatest amount of improvement in the regression equation occurred at Step 2. The F-ratio improved to 149.6, the standard error of the estimate dropped to 0.0057, and the multiple correlation coefficient improved to .94. From this information, it seems that a regression equation with wind speed and APCD ozone as the only independent variables may constitute a sufficient equation. These variables also make sense intuitively. Therefore, the stepwise process will be repeated and forced to terminate at the end of Step 2. However, a residual analysis will be done first so that the residuals generated by the current equation can be compared to the residuals obtained in future trials (Problems #2 and #3).

To initiate the residual analysis type YES when asked. The computer will print the problem number, the dependent variable, and the independent variables that are in the regression equation in order to help identify the data. Note, however, that the regression equation is not printed. Be sure to keep all of the following sheets (residual and plots) with the "Steps" sheets. The computer now prints the number of the observation, the actual measured value of the dependent variable (labeled "ACT."), the value of the dependent variable calculated from the regression equation (labeled "CALC."), and the unit normal deviate (labeled NORM.DEV). On the next page, a histogram of the unit normal deviates is plotted. This plot should resemble the "normal bell-shaped curve", centered on zero. This is a visual test of the assumption of normality. In this example, the curve reasonably approximates a normal distribution. At the bottom of the page is a test for extreme values (outliners). The critical value is printed when the number of observations is 30 or less. It is not

```
CONTROL DATA FOR PROBLEM NO.  2

F-LEVEL FOR INCLUSION (0=.01)?0
F-LEVEL FOR DELETION (0=.005)?0
TOLERANCE LEVEL (0=.001)?0

VARIABLE CONROL VALUES
   0 - DELETE VARIABLE FROM ANALYSIS
   1 - DEPENDENT VARIABLE
   2 - FREE VARIABLE - MAY BE USED IN ANALYSIS
   3 TO 9 - FORCED VARIABLE - LOW TO HIGH LEVEL

CONTROL VALUE FOR:
TIME/100  1 = ?0
 VAN OZ.  2 = ?1
APCD OZ.  3 = ?2
SKY CODE  4 = ?0
 INV/100  5 = ?0
WIND SP.  6 = ?2

THIS PROBLEM MAY REQUIRE UP TO 4 STEPS TO SOLVE.

ENTER THE MAXIMUM NUMBER OF STEPS DESIRED FOR SOLUTION?4
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER          2                       F TO ENTER          .01000
STEP NUMBER             1                       F TO REMOVE         .00500
DEPENDENT VARIABLE      2 (VAN OZ.)             TOLERANCE LEVEL    .00100

VARIABLE ENTERED        6 (WIND SP.)
MULT. CORR. COEFF.       .87828
STD. ERROR EST.          .00791


                                ANOVA TABLE

                    DF        SUM OF SQ.         MEAN SQ.          F-RATIO

REGRESSION           1          .00843            .00843         134.96144
RESIDUAL            40          .00250            .00006
TOTAL               41          .01093



                            VARIABLES IN REGRESSION

 VARIABLE     COEFFICIENT            STD.   COMPUTED     BETA     F-OUT   TYP
                                    ERROR   T- VALUE    COEFF.

 INTERCEPT       .00052
WIND SP.  6      .00843            .00073    11.617      .878   1.3E+02(2)


                         VARIABLES NOT IN REGRESSION

 VARIABLE       PART. CORR.  TOLERANCE     F-IN    TYP

APCD OZ.  3        .70396        .534     3.8E+01(2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER          2                          F TO ENTER          .01000
STEP NUMBER             2                          F TO REMOVE         .00500
DEPENDENT VARIABLE      2 (VAN OZ.)                TOLERANCE LEVEL     .00100

VARIABLE ENTERED        3 (APCD OZ.)
MULT. CORR. COEFF.       .94057
STD. ERROR EST.          .00569


                              ANOVA TABLE

                  DF       SUM OF SQ.          MEAN SQ.          F-RATIO

REGRESSION         2         .00967             .00484          149.58660
RESIDUAL          39         .00126             .00003
TOTAL             41         .01093


                        VARIABLES IN REGRESSION

VARIABLE       COEFFICIENT          STD.   COMPUTED     BETA     F-OUT   TYP
                                   ERROR   T- VALUE    COEFF.

 INTERCEPT        -.00256
APCD OZ.    3      .49114          .07935     6.190      .461   3.8E+01(2)
WIND SP.    6      .00542          .00071     7.581      .564   5.7E+01(2)


F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION.
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


                                    SUMMARY TABLE

 STEP           VARIABLE          MULT.        STD.            F         NO. IN
NUMBER       NAME   IN  OUT       CORR.        ERROR         RATIO        REG.

   1     WIND SP.    6           .87828       .00791     134.96144        1
   2     APCD OZ.    3           .94057       .00569     149.58660        2



RESIDUAL ANALYSIS (YES OR NO)?YES
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 2                    DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   3-APCD OZ.  ,   6-WIND SP.

  OBS.          ACT.          CALC.        RESIDUAL       NORM.DEV

    1         .03000        .02791         .00209          .36710
    2         .03000        .02791         .00209          .36710
    3         .03000        .02842         .00158          .27839
    4         .05000        .04416         .00584         1.02704
   5         .05000        .04416         .00584         1.02704
    6         .05000        .04958         .00042          .07455
    7         .05000        .05990        -.00990        -1.74172
    8         .07000        .06481         .00519          .91192
    9         .07000        .06481         .00519          .91192
   10         .06000        .05940         .00060          .10570

   11         .05000        .05398        -.00398         -.70051
   12         .03000        .03874        -.00874        -1.53788
   13         .03000        .03874        -.00874        -1.53788
   14         .02000        .02300        -.00300         -.52783
   15         .01000        .01809        -.00809        -1.42276
   16         .02000        .01809         .00191          .33595
   17         .02000        .01859         .00141          .24724
   18         .02000        .02351        -.00351         -.61654
   19         .03000        .02842         .00158          .27839
   20         .03000        .02351         .00649         1.14217

   21         .04000        .03434         .00566          .99589
   22         .05000        .04416         .00584         1.02704
   23         .07000        .06481         .00519          .91192
   24         .06000        .06973        -.00973        -1.71057
   25         .06000        .04958         .01042         1.83326
   26         .04000        .03383         .00617         1.08460
   27         .03000        .02892         .00108          .18967
   28         .03000        .02351         .00649         1.14217
   29         .02000        .01809         .00191          .33595
   30         .02000        .01809         .00191          .33595

   31         .03000        .02351         .00649         1.14217
   32         .03000        .02892         .00108          .18967
   33         .04000        .04466        -.00466         -.82038
   34         .05000        .04466         .00534          .93833
   35         .04000        .04466        -.00466         -.82038
   36         .04000        .04466        -.00466         -.82038
   37         .04000        .03925         .00075          .13211
   38         .03000        .03925        -.00925        -1.62660
   39         .03000        .03434        -.00434         -.76282
   40         .02000        .02351        -.00351         -.61654
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 2                       DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   3-APCD OZ.  ,   6-WIND SP.

  OBS.          ACT.          CALC.         RESIDUAL       NORM.DEV

   41          .01000        .01859         -.00859       -1.51147
   42          .01000        .01318         -.00318        -.55898
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 2                        DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   3-APCD OZ.  ,   6-WIND SP.

     HISTOGRAM OF RESIDUALS (NORMAL DEVIATES)

    :***
    :****
    :
 -1 +
    :*****
    :****
    :
  0 +**
    :***********
    :
    :
  1 +*********
    :***
    :
    :*


      TEST FOR EXTREME RESIDUAL VALUES  (DIXON CRITERION)

          R22 (MIN) = .040           R22 (MAX) = .200

         R(7)  =      -.010         R(25)  =       .010
        R(24)  =      -.010         R(31)  =       .006
        R(38)  =      -.009         R(28)  =       .006



   DO YOU WISH PLOTS (YES OR NO)?NO
ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)?YES
```

valid when the critical value is not printed. If the printed values are less than the critical value then there are no outliers.

Since it appears that a shorter regression equation will be valid, plots are not obtained. It is necessary to obtain plots only for the final regression equation.

Problem 2 is the same as problem 1 except that variables 1, 4, and 5 are excluded. This will stop the stepwise process at Step 2. Alternate methods of stopping the process would be to use an F level for inclusion of 2.4 (see Step 2 of Problem 1) or enter a maximum number of 2 steps. The first two steps of problem 2 are exactly the same as problem 1. Residuals are calculated. The resulting histogram of the residuals for this problem not very similiar to a normal distribution.

It was decided at this point to try one more variable in the analysis to see if a better histogram can be obtained. This brings in the variable TIME/100. The computed t-value for this variable is -1.5. The critical value for 41 degrees of freedom is 2.02. It was decided to include this variable in an attempt to improve the histogram on Page 204. However, in addition to violating the t-test, the plot of TIME/100 versus VAN OZONE indicates that another term should be considered in the analysis (a square or cross-product term). Thus, the best regression equation from this set of data is obtained at the end of Step 2.

The plots are examined now in order to see if the equation if adequate. Look for a nonuniform scatter in the plots of residuals versus:

1. observations
2. calculated
3. all independent variables

```
CONTROL DATA FOR PROBLEM NO.  3

F-LEVEL FOR INCLUSION (0=.01)?0
F-LEVEL FOR DELETION (0=.005)?0
TOLERANCE LEVEL (0=.001)?0

VARIABLE CONROL VALUES
   0 - DELETE VARIABLE FROM ANALYSIS
   1 - DEPENDENT VARIABLE
   2 - FREE VARIABLE - MAY BE USED IN ANALYSIS
   3 TO 9 - FORCED VARIABLE - LOW TO HIGH LEVEL

CONTROL VALUE FOR:
TIME/100   1 = ?2
 VAN OZ.   2 = ?1
APCD OZ.   3 = ?2
SKY CODE   4 = ?0
 INV/100   5 = ?0
WIND SP.   6 = ?2

THIS PROBLEM MAY REQUIRE UP TO 6 STEPS TO SOLVE.

ENTER THE MAXIMUM NUMBER OF STEPS DESIRED FOR SOLUTION?6
```

```
STEPWISE MULTIPLE LINEAR REGRESSION


COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


PROBLEM NUMBER             3                        F TO ENTER              .01000
STEP NUMBER                1                        F TO REMOVE             .00500
DEPENDENT VARIABLE         2 (VAN OZ.)              TOLERANCE LEVEL         .00100

VARIABLE ENTERED           6 (WIND SP.)
MULT. CORR. COEFF.          .87828
STD. ERROR EST.             .00791


                                  ANOVA TABLE

                   DF          SUM OF SQ.          MEAN SQ.           F-RATIO

REGRESSION          1            .00843             .00843          134.96144
RESIDUAL           40            .00250             .00006
TOTAL              41            .01093



                             VARIABLES IN REGRESSION

VARIABLE       COEFFICIENT           STD.    COMPUTED     BETA     F-OUT   TYP
                                    ERROR    T- VALUE    COEFF.

  INTERCEPT       .00052
WIND SP.   6      .00843            .00073     11.617       .878   1.3E+02(2)



                        VARIABLES NOT IN REGRESSION

VARIABLE          PART. CORR.   TOLERANCE      F-IN     TYP

TIME/100   1        -.21044          .934      1.8      (2)
APCD OZ.   3         .70396          .534      3.8E+01(2)
```

```
    STEPWISE MULTIPLE LINEAR REGRESSION


    COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


    PROBLEM NUMBER          3                     F TO ENTER          .01000
    STEP NUMBER             2                     F TO REMOVE         .00500
    DEPENDENT VARIABLE      2 (VAN OZ.)           TOLERANCE LEVEL     .00100

    VARIABLE ENTERED        3 (APCD OZ.)
    MULT. CORR. COEFF.       .94057
    STD. ERROR EST.          .00569


                                 ANOVA TABLE

                      DF       SUM OF SQ.        MEAN SQ.          F-RATIO

    REGRESSION         2         .00967           .00484        149.58660
    RESIDUAL          39         .00126           .00003
    TOTAL             41         .01093


                           VARIABLES IN REGRESSION

    VARIABLE    COEFFICIENT            STD.   COMPUTED    BETA     F-OUT   TYP
                                      ERROR   T- VALUE   COEFF.

   INTERCEPT       -.00256
 APCD OZ.   3       .49114          .07935    6.190      .461   3.8E+01(2)
 WIND SP.   6       .00542          .00071    7.581      .564   5.7E+01(2)


                        VARIABLES NOT IN REGRESSION

    VARIABLE        PART. CORR.  TOLERANCE    F-IN    TYP

 TIME/100   1         -.23997         .931    2.3     (2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER            3                        F TO ENTER             .01000
STEP NUMBER               3                        F TO REMOVE            .00500
DEPENDENT VARIABLE        2 (VAN OZ.)              TOLERANCE LEVEL        .00100

VARIABLE ENTERED          1 (TIME/100)
MULT. CORR. COEFF.         .94410
STD. ERROR EST.            .00559

                                 ANOVA TABLE

                    DF        SUM OF SQ.          MEAN SQ.          F-RATIO

REGRESSION           3          .00975             .00325         103.87860
RESIDUAL            38          .00119             .00003
TOTAL               41          .01093


                              VARIABLES IN REGRESSION

 VARIABLE       COEFFICIENT          STD.    COMPUTED     BETA      F-OUT   TYP
                                    ERROR    T- VALUE    COEFF.

  INTERCEPT         .00081
TIME/100   1       -.00034         .00022     -1.524      -.084     2.3      (2)
APCD OZ.   3        .48433         .07816      6.196       .454     3.8E+01(2)
WIND SP.   6        .00567         .00072      7.853       .590     6.2E+01(2)


F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION.
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA


                                 SUMMARY TABLE

STEP          VARIABLE          MULT.        STD.           F          NO. IN
NUMBER     NAME    IN  OUT      CORR.        ERROR        RATIO         REG.

  1    WIND SP.    6           .87828       .00791    134.96144         1
  2    APCD OZ.    3           .94057       .00569    149.58660         2
  3    TIME/100    1           .94410       .00559    103.87860         3



RESIDUAL ANALYSIS (YES OR NO)?YES
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                        DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES     1-TIME/100   ,   3-APCD OZ.   ,   6-WIND SP.

 OBS.        ACT.           CALC.         RESIDUAL        NORM.DEV

   1        .03000         .02949          .00051          .09155
   2        .03000         .02915          .00085          .15199
   3        .03000         .02963          .00037          .06534
   4        .05000         .04547          .00453          .80980
   5        .05000         .04513          .00487          .87025
   6        .05000         .05046         -.0006          -.08251
   7        .05000         .06063         -.01063        -1.90137
   8        .07000         .06514          .00486          .86950
   9        .07000         .06480          .00520          .92995
  10        .06000         .05880          .00120          .21534

  11        .05000         .05279         -.00279         -.49927
  12        .03000         .03710         -.00710        -1.26994
  13        .03000         .03676         -.00676        -1.20950
  14        .02000         .02025         -.00025         -.04480
  15        .01000         .01980         -.00980        -1.75277
  16        .02000         .01946          .00054          .09594
  17        .02000         .01995          .00005          .00928
  18        .02000         .02445         -.00445         -.79637
  19        .03000         .02896          .00104          .18623
  20        .03000         .02378          .00622         1.11279

  21        .04000         .03477          .00523          .93509
  22        .05000         .04412          .00588         1.05159
  23        .07000         .06480          .00520          .92995
  24        .06000         .06930         -.00930        -1.66397
  25        .06000         .04877          .01123         2.00799
  26        .04000         .03226          .00774         1.38443
  27        .03000         .02708          .00292          .52271
  28        .03000         .02107          .00893         1.59637
  29        .02000         .01980          .00020          .03549
  30        .02000         .01946          .00054          .09594

  31        .03000         .02479          .00521          .93144
  32        .03000         .03012         -.00012         -.02131
  33        .04000         .04596         -.00596        -1.06512
  34        .05000         .04562          .00438          .78359
  35        .04000         .04528         -.00528         -.94422
  36        .04000         .04494         -.00494         -.88378
  37        .04000         .03894          .00106          .18988
  38        .03000         .03860         -.00860        -1.53794
  39        .03000         .03342         -.00342         -.61139
  40        .02000         .02175         -.00175         -.31279
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                      DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.

  OBS.        ACT.          CALC.        RESIDUAL      NORM.DEV

   41        .01000        .01657        -.00657      -1.17451
   42        .01000        .01056        -.00056       -.10085
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                    DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.

      HISTOGRAM OF RESIDUALS (NORMAL DEVIATES)

  -2 +*
     :**
     :*
     :***
  -1 +***
     :*
     :**
     :*
   0 +**********
     :****
     :*
     :****
   1 +******
     :
     :**
     :
   2 +*


        TEST FOR EXTREME RESIDUAL VALUES  (DIXON CRITERION)

           R22 (MIN) = .072           R22 (MAX) = .170

          R(7)   =      -.011        R(25)   =       .011
         R(15)   =      -.010        R(28)   =       .009
         R(24)   =      -.009        R(26)   =       .008


DO YOU WISH PLOTS (YES OR NO)?YES
```

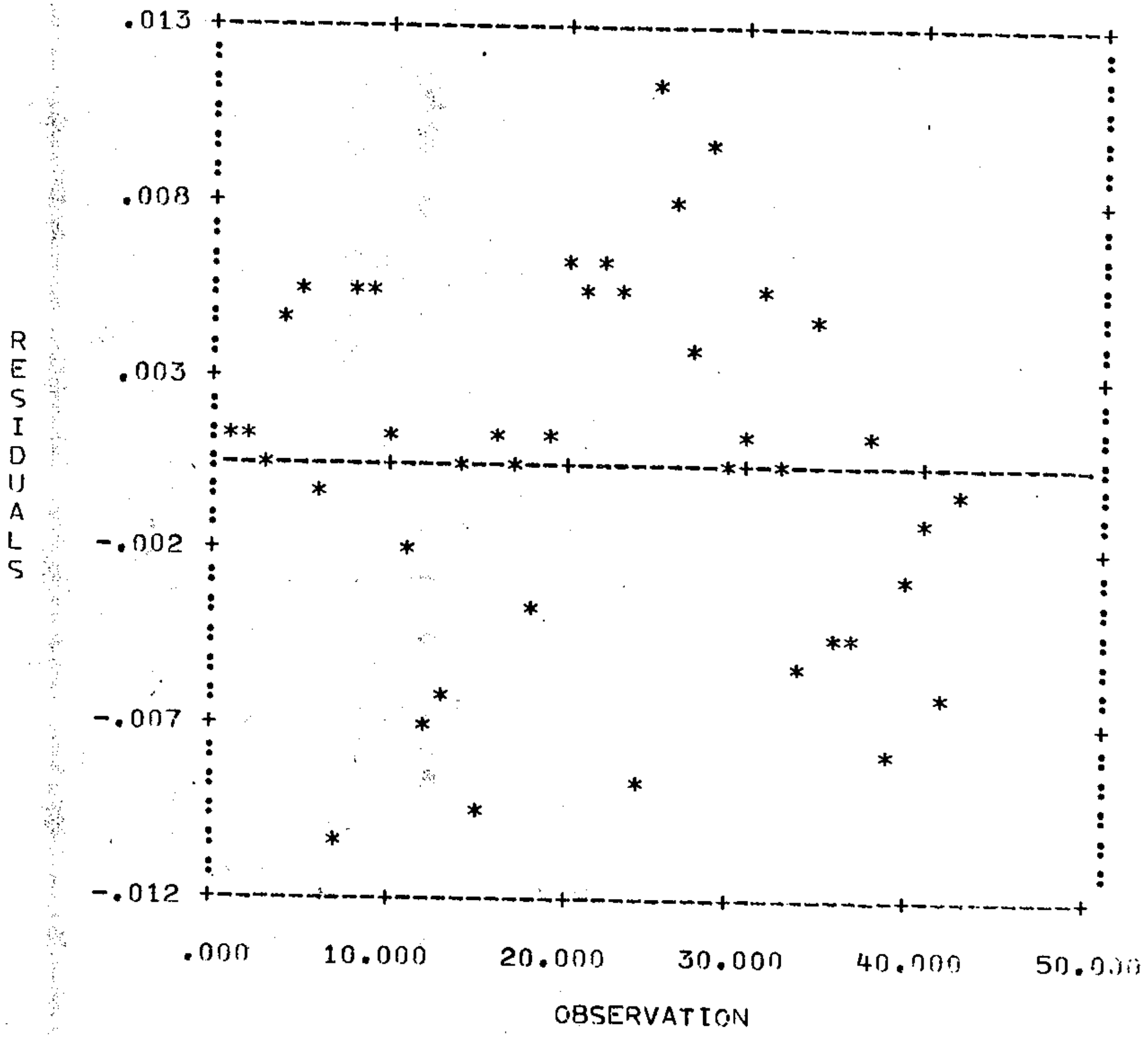
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3   DEPENDENT VARIABLE 2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100 , 3-APCD OZ. , 6-WIND SP.

OBSERVATIONS VS RESIDUALS

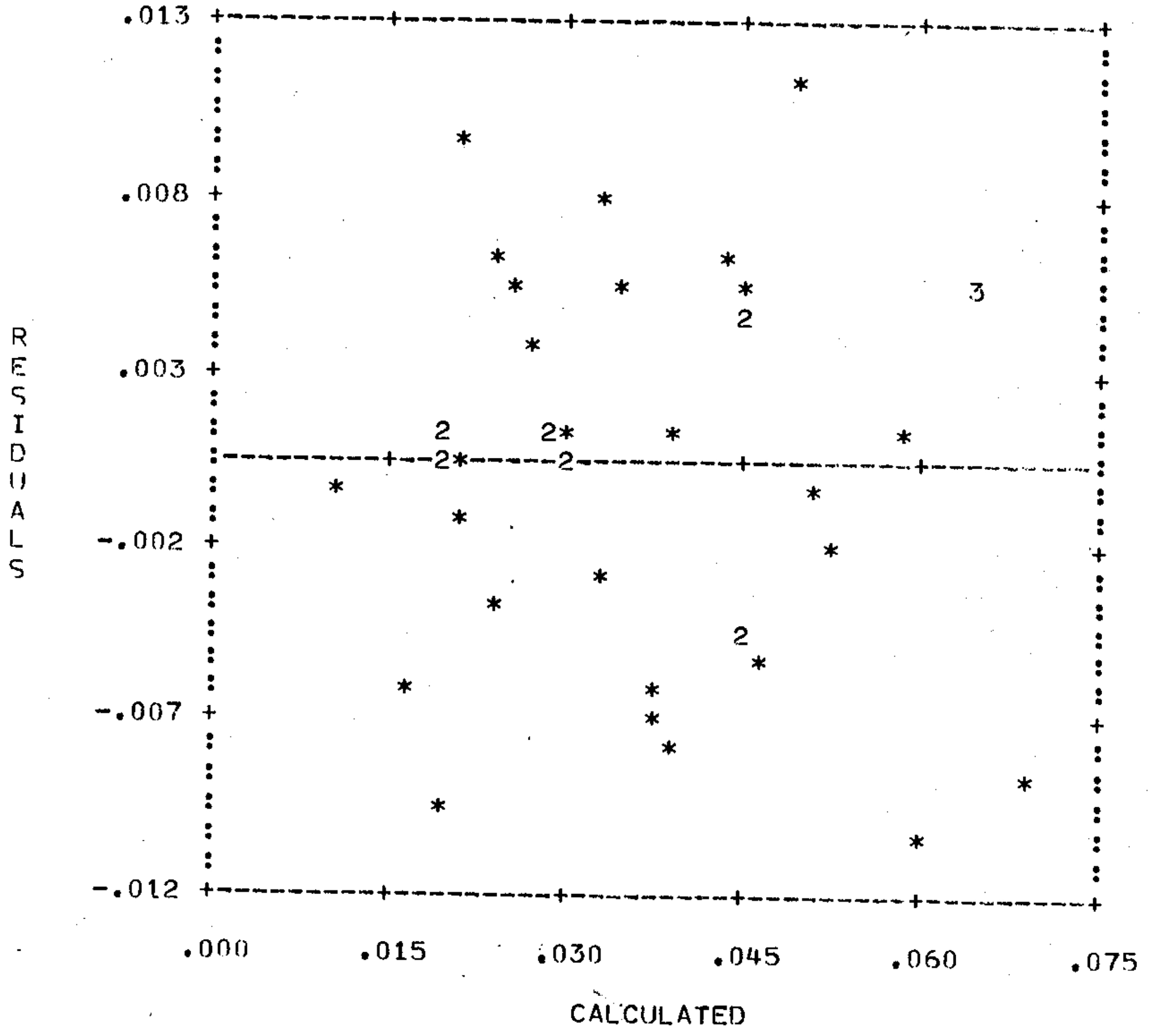


STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

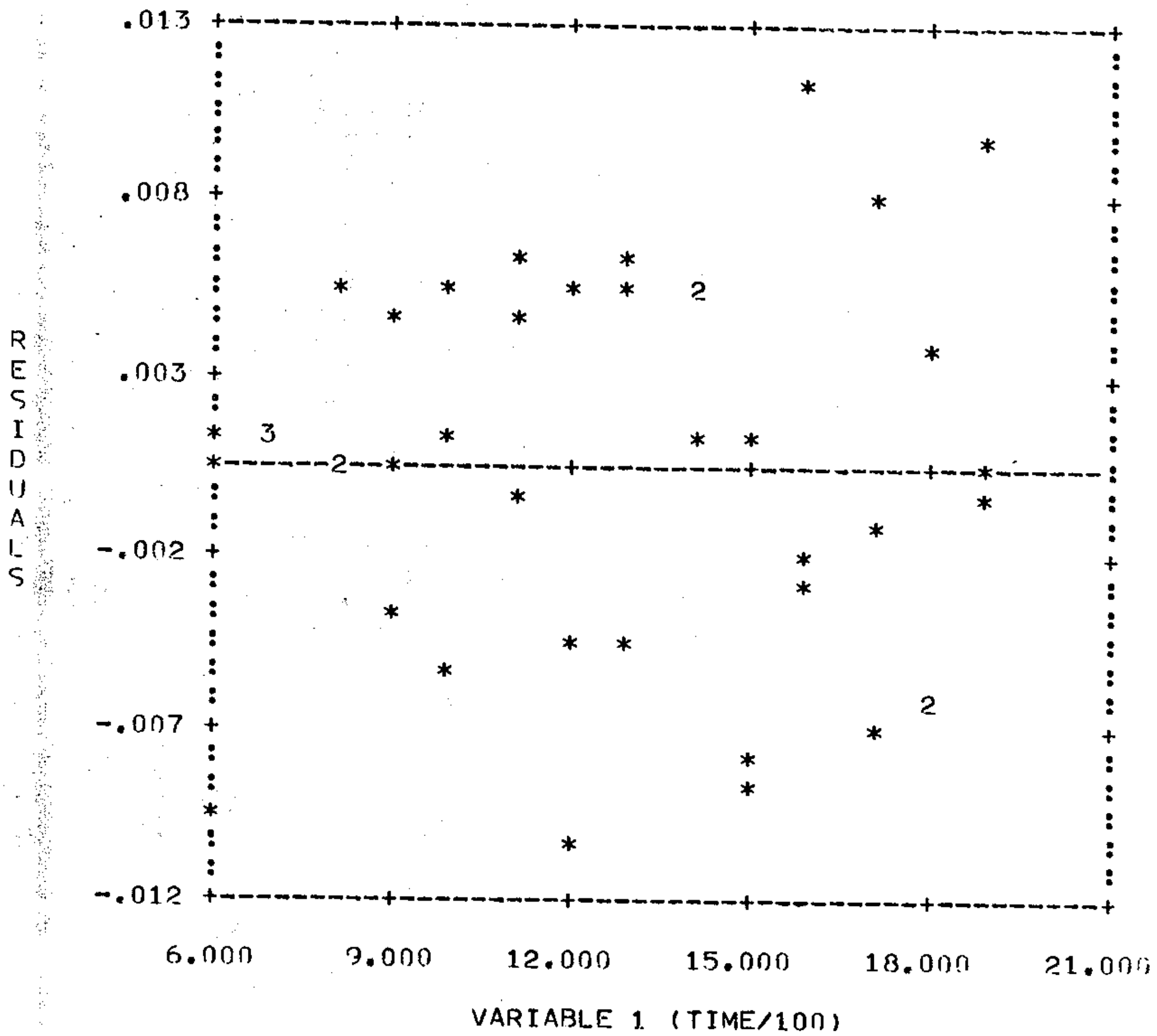
```
PROBLEM NUMBER 3                          DEPENDENT VARIABLE  2-VAN OZ.
INDEPEND. VARIABLES    1-TIME/100   ,    3-APCD OZ.   ,    6-WIND SP.
```

CALCULATED VS RESIDUALS



```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                        DEPENDENT VARIABLE  2-VA OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.
```

VARIABLE 1 (TIME/100) VS RESIDUALS



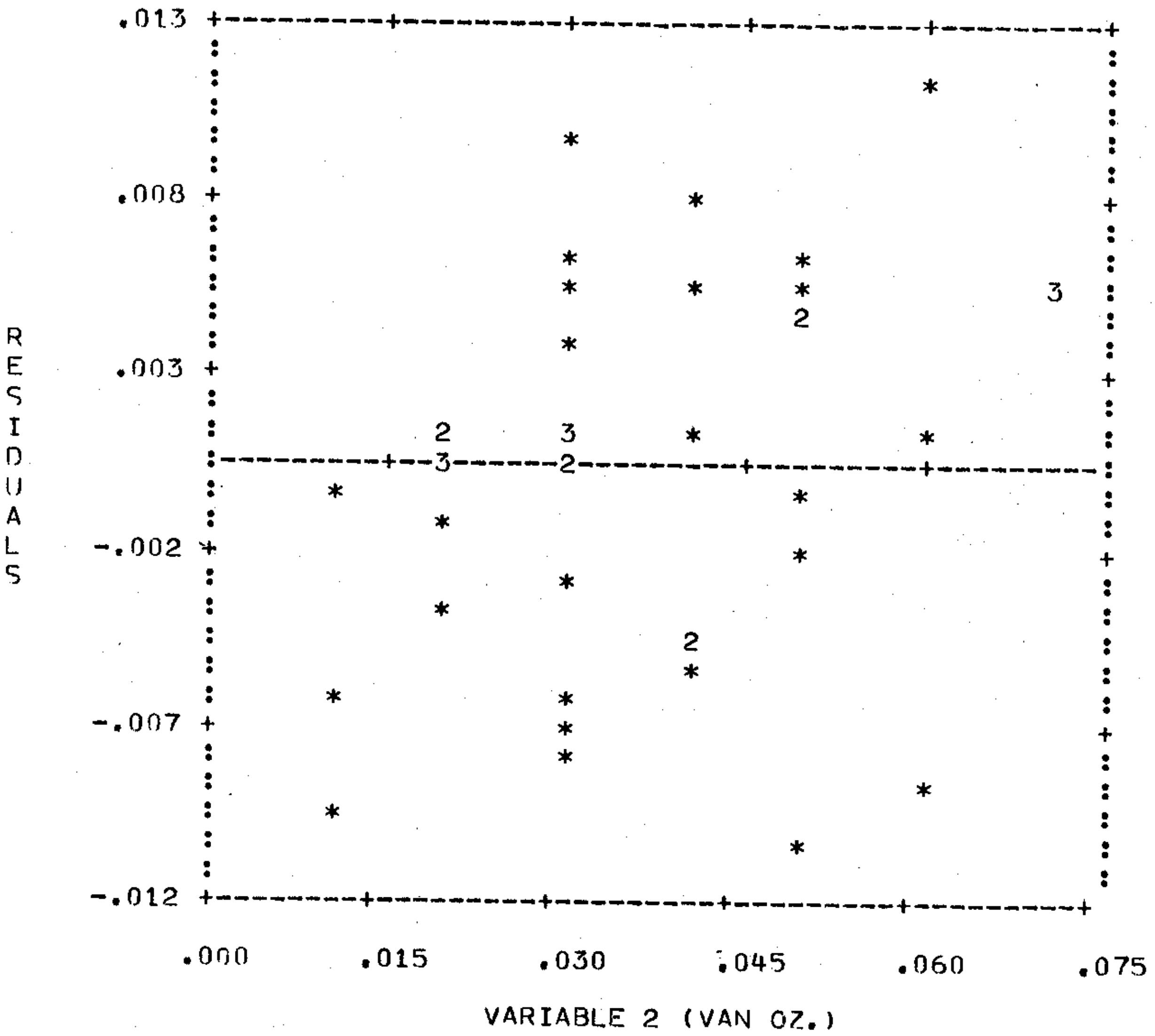
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

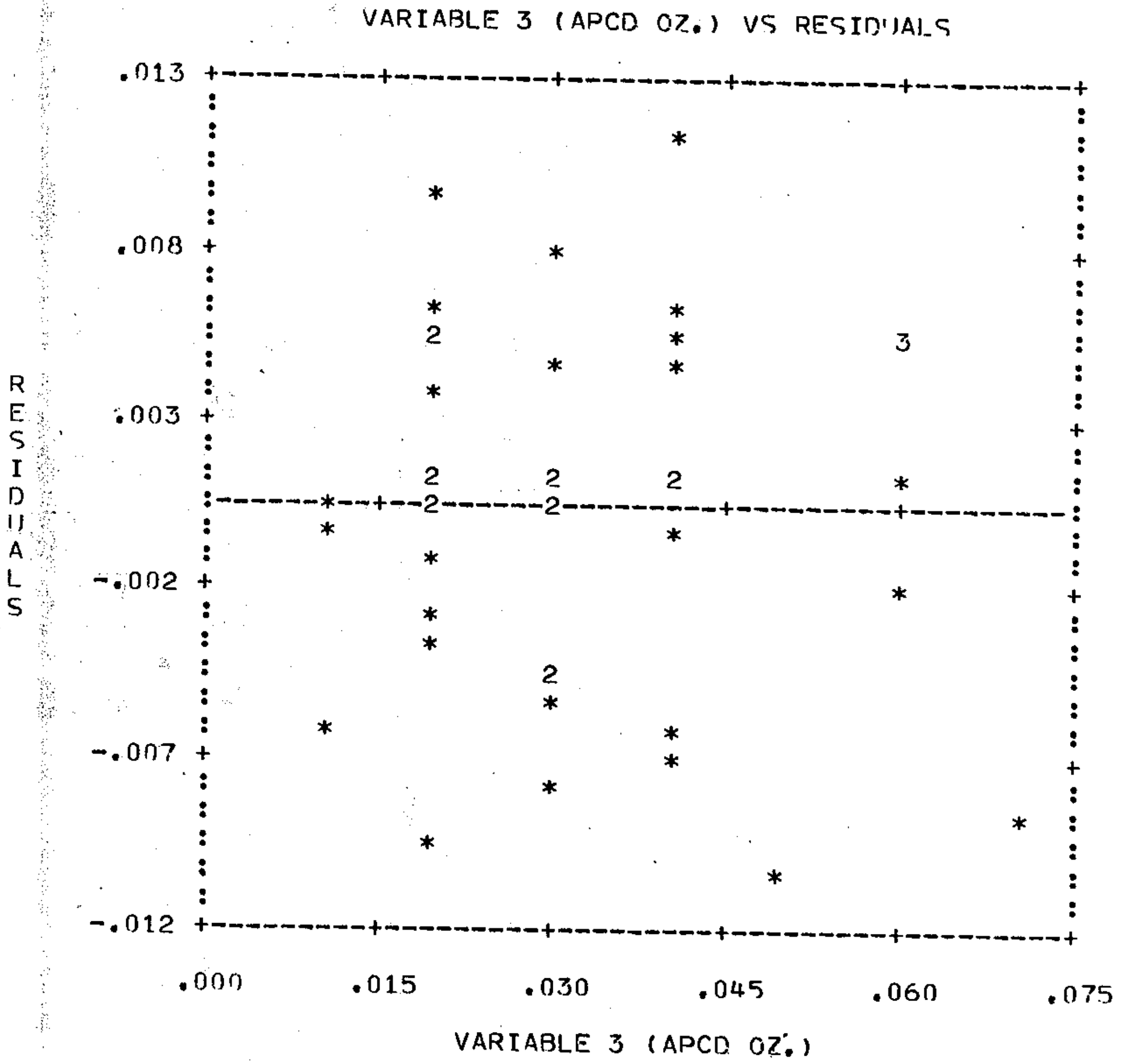
PROBLEM NUMBER 3 DEPENDENT VARIABLE 2-VAN OZ.

INDEPEND. VARIABLES 1-TIME/100 , 3-APCD OZ. , 6-WIND SP.

VARIABLE 2 (VAN OZ.) VS RESIDUALS

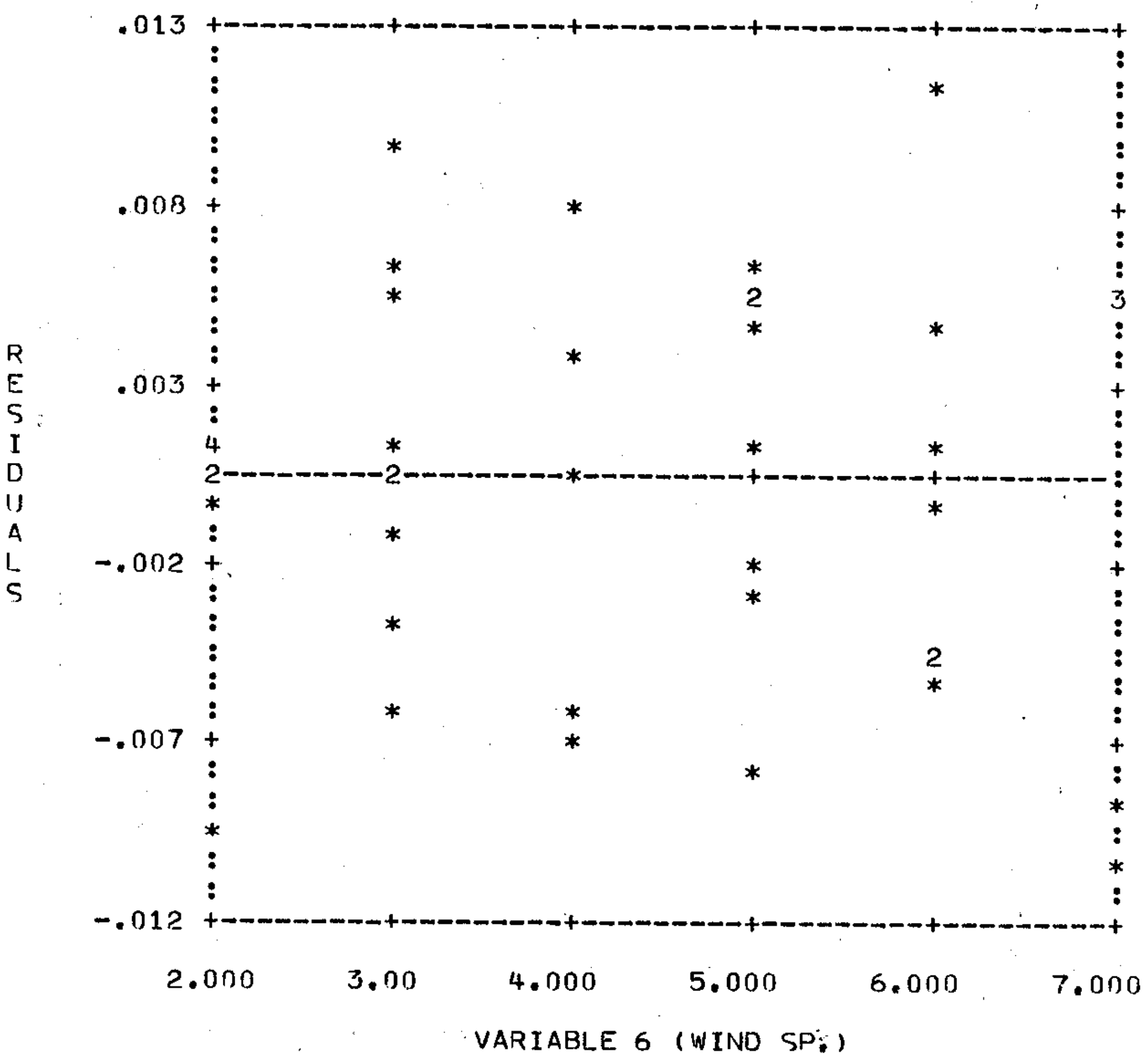


```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                        DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.
```



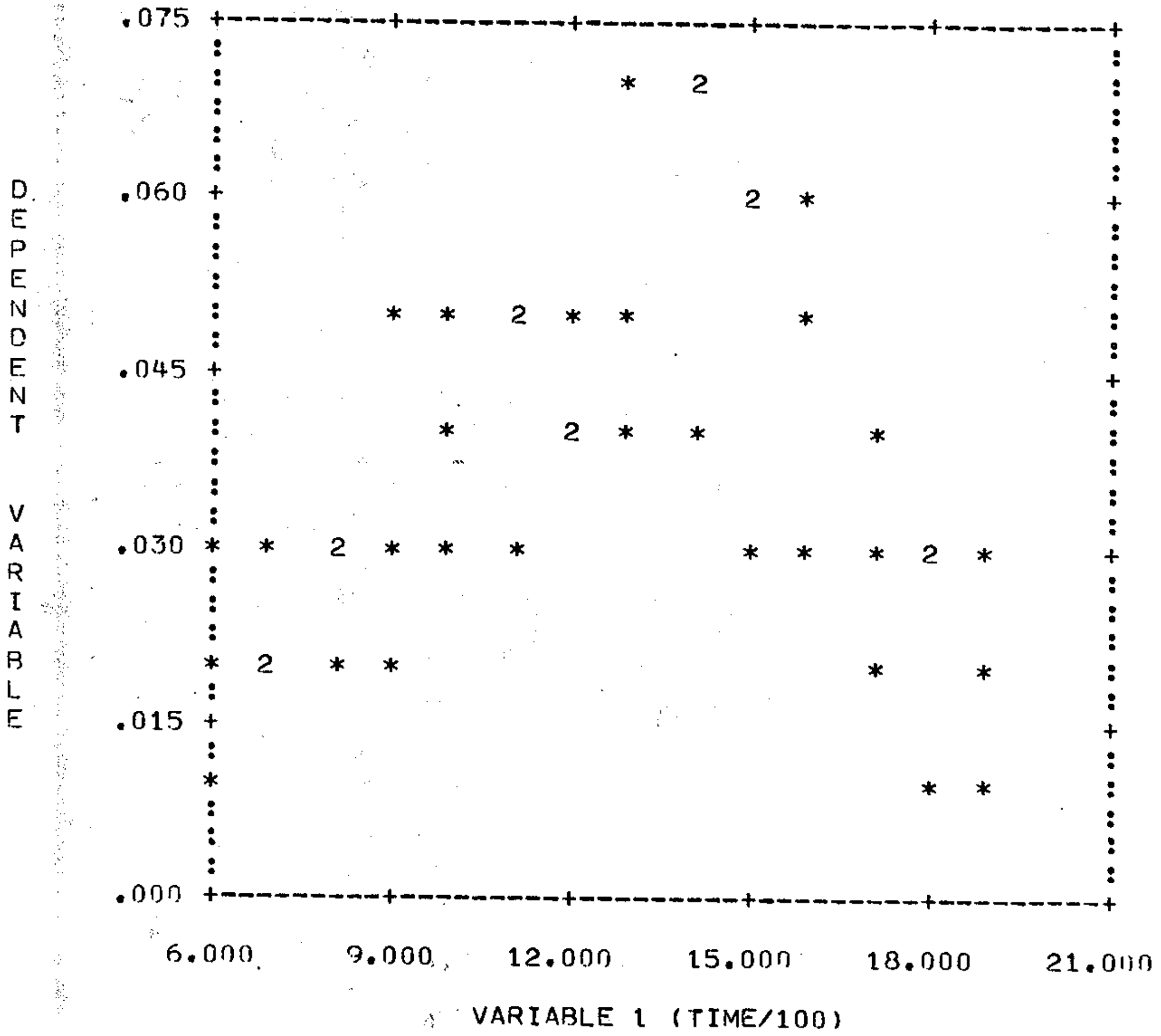
```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                        DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.
```

VARIABLE 6 (WIND SP.) VS RESIDUALS



```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3                          DEPENDENT VARIABLE  2-VAN OZ.

INDEPEND. VARIABLES   1-TIME/100  ,   3-APCD OZ.  ,   6-WIND SP.
```

VARIABLE 1 (TIME/100) VS VARIABLE 2 (VAN OZ.)



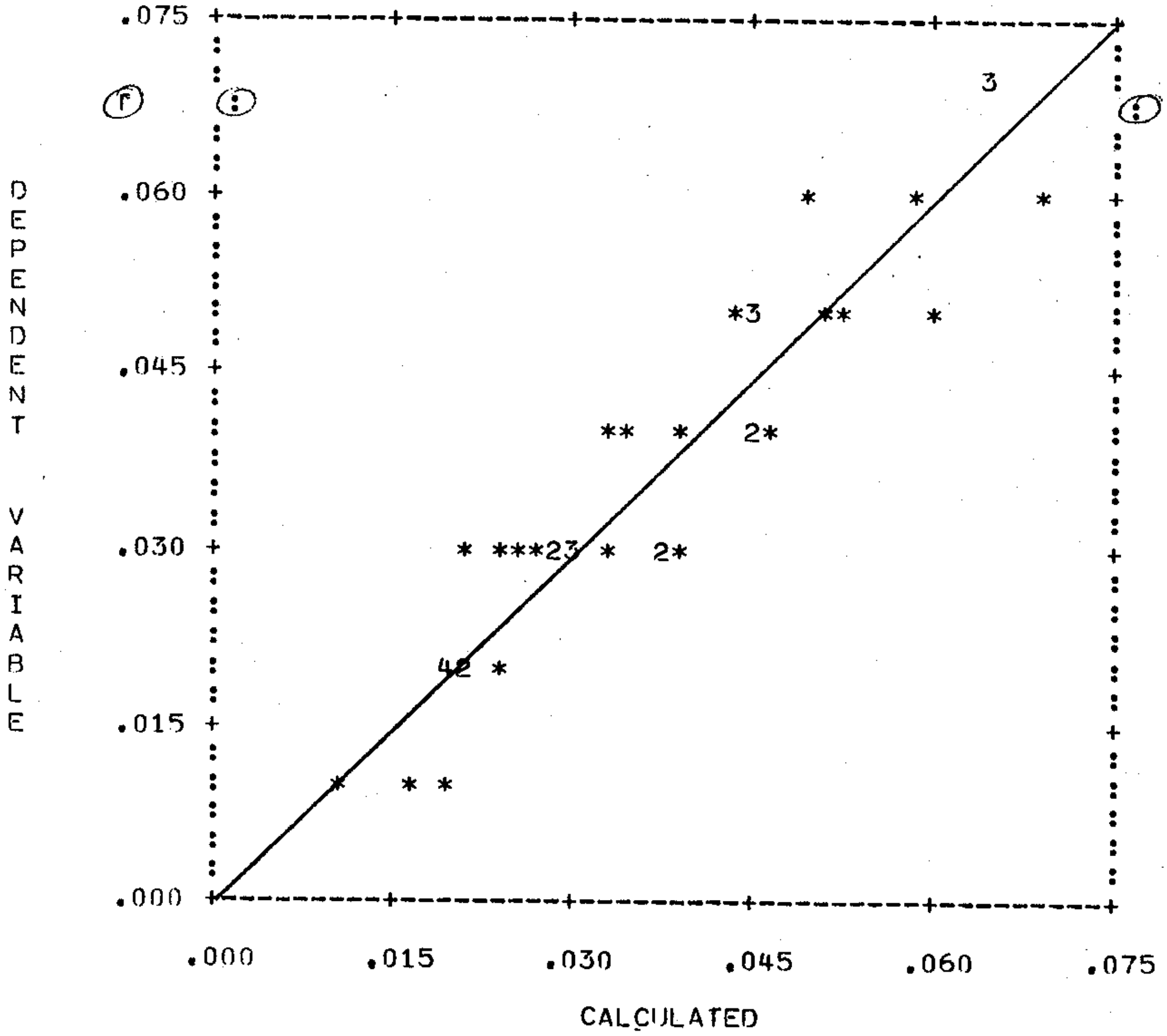
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3 DEPENDENT VARIABLE 2-VAN OZ.

INDEPEND. VARIABLES 1-TIME/100 , 3-APCD OZ. , 6-WIND SP.

CALCULATED VS VARIABLE 2 (VAN OZ.)



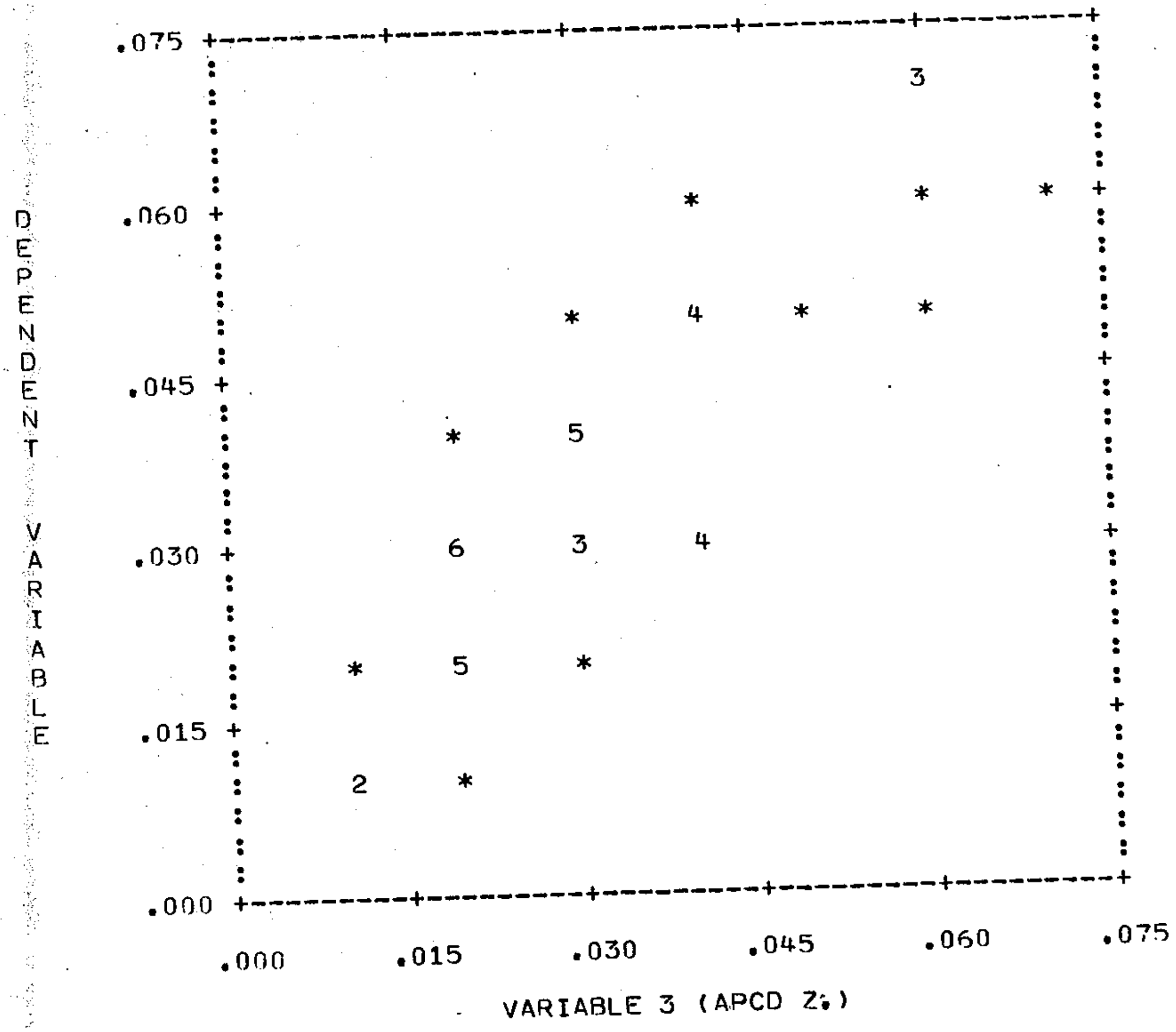
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3 DEPENDENT VARIABLE 2-VAN OZ.

INDEPEND. VARIABLES 1-TIME/100 , 3-APCD OZ. , 6-WIND SP.

VARIABLE 3 (APCD OZ.) VS VARIABLE 2 (VAN OZ.)

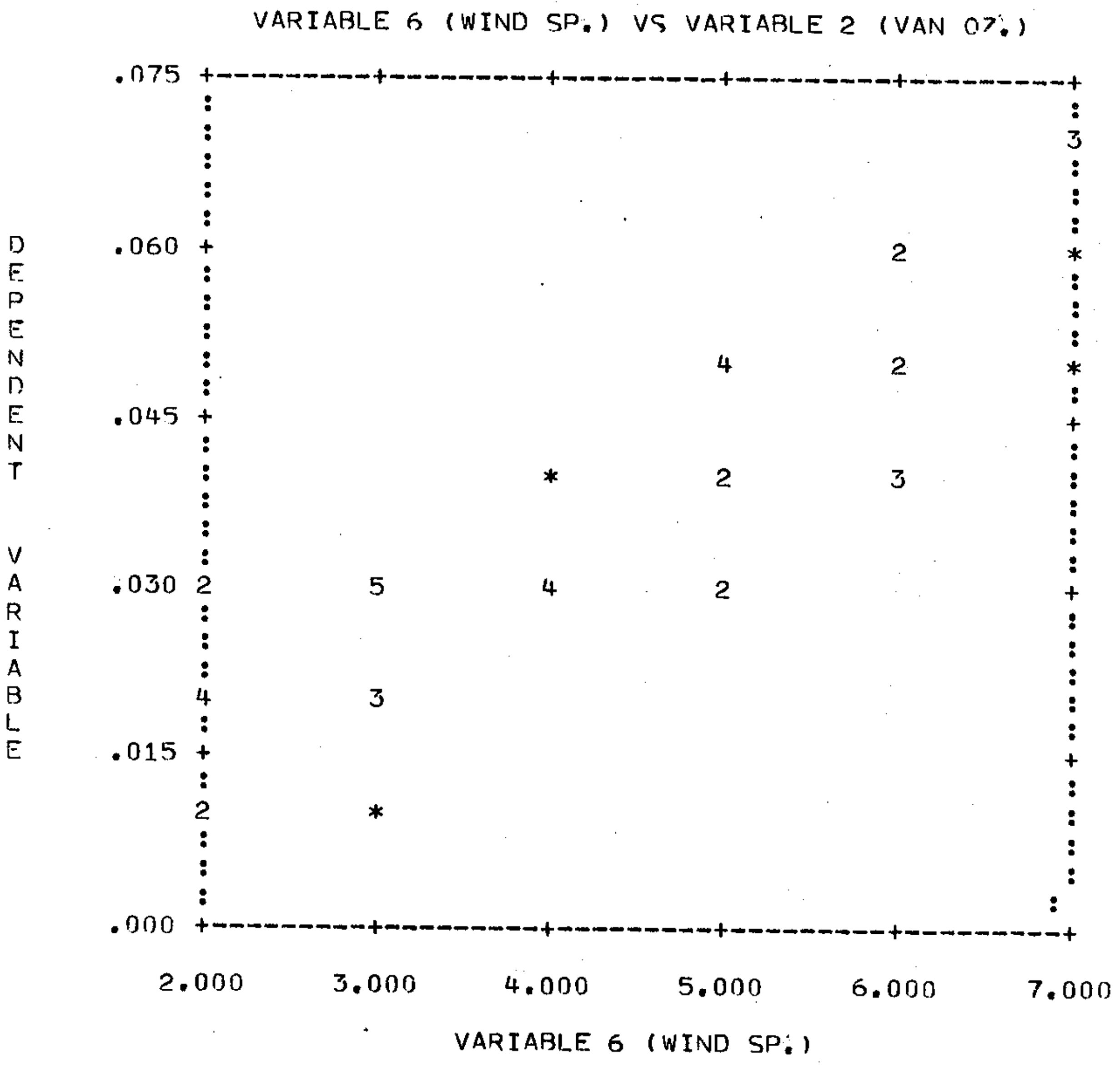


STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE VAN OZONE TO APCD OZONE--SAMPLE DATA

PROBLEM NUMBER 3 DEPENDENT VARIABLE 2-VAN OZ.

INDEPEND. VARIABLES 1-TIME/100 , 3-APCD OZ. , 6-WIND SP.



```
    ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)?NO
    INPUT NEW DATA FILENAME OR 'STOP'?STOP

       GOOD-BYE.  BE SURE TO DELETE FILE '$REG' BEFORE LEAVING
    THE SYSTEM.
>QUIT
-DEL $REG
-LOGO
 1342  12/26/73
CPU MINS -  1.840
TERMINAL MINS -  47.50
FILE MODULES - 37
----------------






       TENET 210 TIME-SHARING SYSTEM  1342  12/26/73 39

-LOGIN
```

Next, the dependent variable is plotted versus each independent variable. The plot of calculated versus dependent should be a straight line through the origin with a slope of 1.

At the end of the run, note the steps required to complete the run and delete $REG.

A computer program facilitates use of the regression equation to fill in the missing data points. The best derived regression equation is (from Step 2, Problem 2):

$$\text{Van ozone} = -2.56\ 10^{-3} + .49\ (\text{APCD}) + 5.42 \times 10^{-3}\ (\text{Wind Speed})$$

This can be incorporated into a simple computer program. A datafile is created using the regression equation and then the data is merged with the original data set to obtain the complete data set.

The simple program written for the above regression equation is:

```
100 OPEN 'STEP1', 1,INPUT,OLD
110 OPEN 'STEP3',2
120 ON ENDFILE(1) GOTO 170
130 INPUT FROM 1:TME,APCD,IINV,SKY,U
140 OZ=2.56E-3+ .49(APCD)+5.42E-3*(U)
150 PRINT ON 2 IN FORM "2% 2 (2%.2%) 3% 4% 3%": TME,OZ,APCD,SKY,IINV,U
160 GOTO 10
170 END
```

Once this program has been run and the data checked, the easiest thing to do is to go to the executive mode (-) and append STEP3 to STEP as follows:

```
-APPEND STEP3 TO STEP
```

This will place the data from STEP3 at the end of STEP in the STEP datafile. Then STEP3 can be deleted.

A SIMPLE COMPUTER PROGRAM

The user may wish to create a small data file from a large datafile. A simple computer program may be created by the user for this purpose. For example, suppose that it is desired to use just two of the variables in the STEP datafile. The following is a listing of an example computer program which could be used.

This program is <u>not</u> a "canned" program.

```
100 OPEN 'STEP',1,INPUT,OLD
110 OPEN 'STEP9',2
120 ON ENDFILE(1) GOTO 170
130 INPUT FROM 1:X1,X2,X3,X4,X5,X6
140 IF X1=-1 OR X2=-1 OR X3=-1 OR X4=-1 OR X5=-1 OR X6=-1 GOTO 130
150 PRINT ON 2:X2,X3
160 GOTO 130
170 END
```

This program will only consider the data which is not -1. It will generate a new datafile (STEP9) which contains only variables number 2 and 3.

Another simple program is shown next. This program reads a datafile, STEP, and creates a new datafile, STEP1, which has the log of the field data in place of the actual concentration. This process is repeated with STEP1 the "old" datafile and STEP2 the new datafile. The log transformation was performed on the APCD ozone values in this step. STEP2 is now converted to a log-log form.

```
>SAVE OLD 'LOG%TRAN'
>LIST
100 PRINT "WHICH DATAFILE DO YOU WISH TO CONVERT TO LOG POLLUTANT VALUES?":
110 INPUT FIL$
120 PRINT
130 PRINT "WHAT DO YOU WISH TO CALL THE NEW DATA FILE":
140 INPUT FL$
150 PRINT
160 PRINT "HOW MANY COLUMNS IN THE FIRST DATAFILE":
170 INPUT N1
180 PRINT
190 PRINT "WHICH COLUMN CONTAINS THE POLLUTANT DATA (INPUT A NUMBER
FROM 1 TO ":N1:")":
200 INPUT N2
210 IF N2<(N1+1) THEN 260
220 PRINT
230 PRINT "COLUMN NUMBER OF THE POLLUTANT DATA MUST BE LESS THAN OR = ":N1
240 PRINT
250 GOTO 190
260 PRINT
270 OPEN FIL$,1,INPUT,OLD
280 OPEN FL$,2
290 ON ENDFILE (1) GOTO 470
300 N3=0
310 FOR I=1 TO N1
320    INPUT FROM 1:X(I)
330    NEXT I
340 IF X(N2)=0 THEN N3=N3+1
350 IF X(N2)=0 THEN 310
360 FOR I=1 TO (N2-1)
370       PRINT ON 2:X(I);
380       NEXT I
390 PRINT ON 2:LOG10(X(N2));
400 FOR I=(N2+1) TO N1
410 PRINT ON 2:X(I);
420 NEXT I
430 PRINT ON 2:""
440 GOTO 310
450 IF N3#0 THEN 460 ELSE 470
460 PRINT N3:" LINES OF DATA WERE OMITTED FROM  ":FL$:" TO AVOID DIVIDING
BY ZERO"
470 PRINT "DO YOU WISH TO CONVERT ANOTHER FILE? (ENTER YES OR NO)":
480 INPUT ANS$
490 PRINT
500 CLOSE 1
510 CLOSE 2
520 IF ANS$='YES' THEN 100
530 END
>
```

```
RUN
WHICH DATAFILE DO YOU WISH TO CONVERT TO LOG POLLUTANT VALUES??452;STAR;STEP
WHAT DO YOU WISH TO CALL THE NEW DATA FILE?STEP1
HOW MANY COLUMNS IN THE FIRST DATAFILE?6
WHICH COLUMN CONTAINS THE POLLUTANT DATA (INPUT A NUMBER FROM 1 TO 6)?2
DO YOU WISH TO CONVERT ANOTHER FILE? (ENTER YES OR NO)?YES
WHICH DATAFILE DO YOU WISH TO CONVERT TO LOG POLLUTANT VALUES??STEP1
WHAT DO YOU WISH TO CALL THE NEW DATA FILE?STEP2
HOW MANY COLUMNS IN THE FIRST DATAFILE?6
WHICH COLUMN CONTAINS THE POLLUTANT DATA (INPUT A NUMBER FROM 1 TO 6)?3
DO YOU WISH TO CONVERT ANOTHER FILE? (ENTER YES OR NO)?NO
>QUIT
```

```
COPY STEP1 TO TEL
01 6    -1.52288    .4E-1   1   5   2
02 7    -1.52288    .4E-1   1   7   2
03 8    -1.52288    .3E-1   2   10   3
04 9    -1.30103    .4E-1   2   12   5
05 10    -1.30103    .4E-1   2   12   5
06 11    -1.30103    .4E-1   2   14   6
07 12    -1.30103    .5E-1   2   16   7
08 13    -1.1549    .6E-1   2   20   7
09 14    -1.1549    .6E-1   2   25   7
10 15    -1.22185    .6E-1   2   25   6
11 16    -1.30103    .6E-1   2   27   5
12 17    -1.52288    .4E-1   3   28   4
13 18    -1.52288    .4E-1   2   30   4
14 19    -1.69897    .3E-1   2   32   2
15 6    -2    .2E-1   0   10   2
16 7    -1.69897    .2E-1   0   10   2
17 8    -1.69897    .1E-1   1   15   3
18 9    -1.69897    .2E-1   1   15   3
19 10    -1.52288    .3E-1   1   16   3
20 11    -1.52288    .2E-1   1   16   3
21 12    -1.39794    .2E-1   1   17   5
22 13    -1.30103    .4E-1   2   20   5
23 14    -1.1549    .6E-1   3   23   7
24 15    -1.22185    .7E-1   3   25   7
25 16    -1.22185    .4E-1   2   25   6
26 17    -1.39794    .3E-1   3   30   4
27 18    -1.52288    .2E-1   3   35   4
28 19    -1.52288    .2E-1   3   35   3
29 6    -1.69897    .2E-1   4   7   2
30 7    -1.69897    .2E-1   4   9   2
31 8    -1.52288    .2E-1   4   10   3
32 9    -1.52288    .2E-1   4   15   4
33 10    -1.39794    .3E-1   5   20   6
34 11    -1.30103    .3E-1   5   25   6
35 12    -1.39794    .3E-1   4   25   6
36 13    -1.39794    .3E-1   4   25   6
37 14    -1.39794    .3E-1   4   27   5
38 15    -1.52288    .3E-1   4   30   5
39 16    -1.52288    .2E-1   2   25   5
40 17    -1.69897    .2E-1   2   24   3
41 18    -2    .1E-1   2   20   3
42 19    -2    .1E-1   1   17   2
-
```

```
COPY STEP2 TO TEL
01 6    -1.52288   -1.39794   1    5    2
02 7    -1.52288   -1.39794   1    7    2
03 8    -1.52288   -1.52288   2    10    3
04 9    -1.30103   -1.39794   2    12    5
05 10    -1.30103   -1.39794   2    12    5
06 11    -1.30103   -1.39794   2    14    6
07 12    -1.30103   -1.30103   2    16    7
08 13    -1.1549   -1.22185   2    20    7
09 14    -1.1549   -1.22185   2    25    7
10 15    -1.22185   -1.22185   2    25    6
11 16    -1.30103   -1.22185   2    27    5
12 17    -1.52288   -1.39794   3    28    4
13 18    -1.52288   -1.39794   2    30    4
14 19    -1.69897   -1.52288   2    32    2
15 6    -2    -1.69897    0    10    2
16 7    -1.69897   -1.69897    0    10    2
17 8    -1.69897   -2    1    15    3
18 9    -1.69897   -1.69897    1    15    3
19 10    -1.52288   -1.52288    1    16    3
20 11    -1.52288   -1.69897    1    16    3
21 12    -1.39794   -1.69897    1    17    5
22 13    -1.30103   -1.39794    2    20    5
23 14    -1.1549   -1.22185    3    23    7
24 15    -1.22185   -1.1549    3    25    7
25 16    -1.22185   -1.39794    2    25    6
26 17    -1.39794   -1.52288    3    30    4
27 18    -1.52288   -1.69897    3    35    4
28 19    -1.52288   -1.69897    3    35    3
29 6    -1.69897   -1.69897    4    7    2
30 7    -1.69897   -1.69897    4    9    2
31 8    -1.52288   -1.69897    4    10    3
32 9    -1.52288   -1.69897    4    15    4
33 10    -1.39794   -1.52288    5    20    6
34 11    -1.30103   -1.52288    5    25    6
35 12    -1.39794   -1.52288    4    25    6
36 13    -1.39794   -1.52288    4    25    6
37 14    -1.39794   -1.52288    4    27    5
38 15    -1.52288   -1.52288    4    30    5
39 16    -1.52288   -1.69897    2    25    5
40 17    -1.69897   -1.69897    2    24    3
41 18    -2    -2    2    20    3
42 19    -2    -2    1    17    2
-
```

STATISTIC SUMMARY PROGRAMS

Descriptive Statistics "MATHISTO"

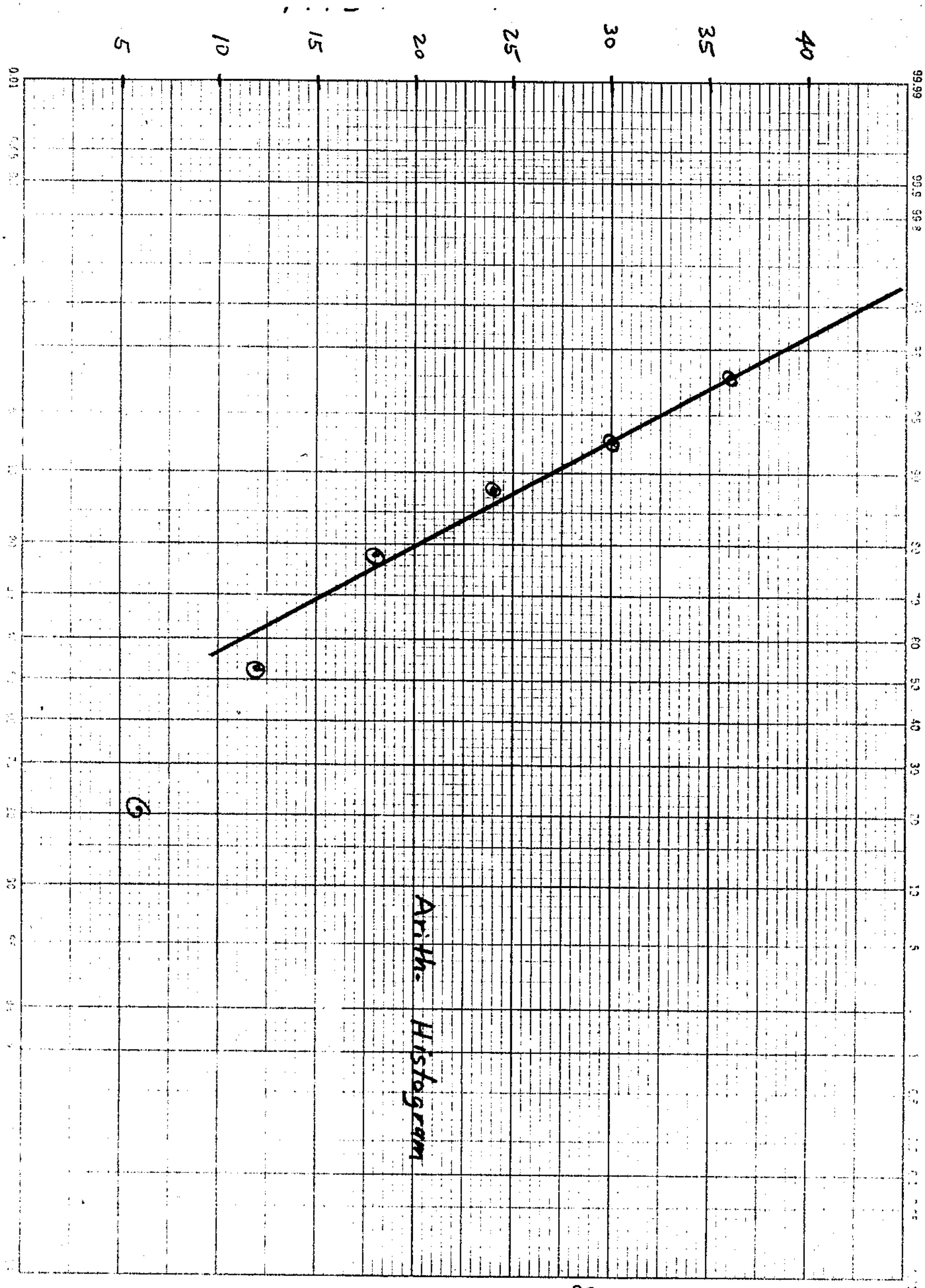
This is a descriptive statistics summary program that will compute X, σ , ranges, coefficient of variation, etc. and histograms for a specified class width interval. From this histogram, a cumulative frequency diagram can be made indicating the % of time a certain concentration or health standard was exceeded.

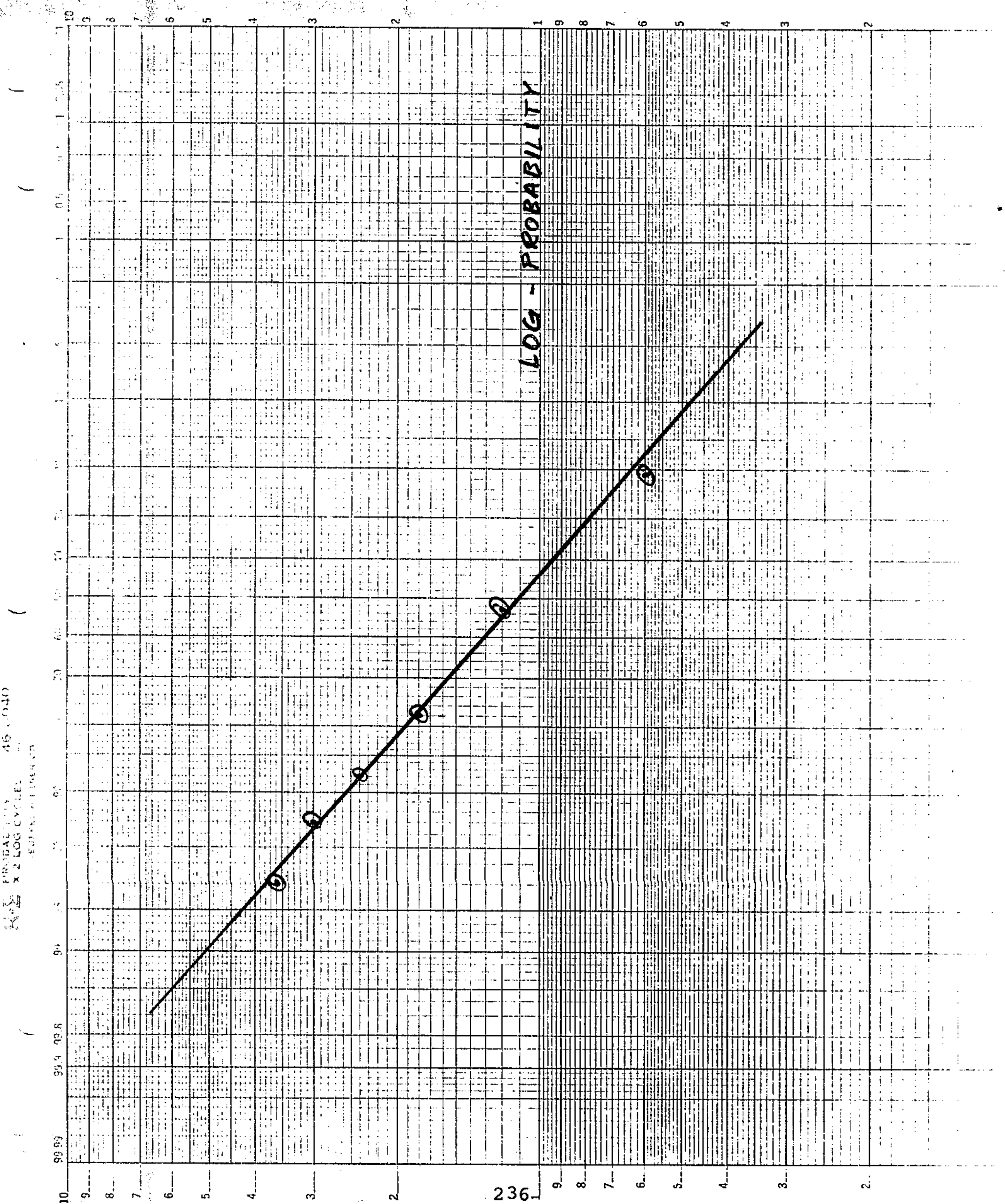
It is recommended that these cumulative frequency plots be made for monthly or seasonal bases. These cumulative frequency diagrams should be made for all pollutants CO, HC, NO_x and Ozone. A trend analysis based on the pollutant burden concept (tons/day) can be used to indicate whether these values will increase or decrease in future years.

Also, from the histogram a visual check can be made to check if the data follows a normal or lognormal curve depending on the analysis. The computer name for the program which produces a histogram of the data is '5;LAB;MATHISTO'.

```
-BASI
>LINK '5,LAB,MATHISTO'
   DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?%FRIED
   JOB TITLE - NO. ROWS -NO. COLUMNS?TEST FRIEDMAN DATA,36,10
   LOG TRANSFORM?  INPUT  1 FOR YES  2 FOR NO ?2
```

```
SAMPLE STATISTICS

TEST FRIEDMAN DATA   PROBE 1

OBSERVATIONS=  3.600E+01    MEAN     =  1.581E+01    STD.DEV.=  7.486E+0
MINIMUM     =  4.000E+00    RANGE    =  3.000E+01    MAXIMUM =  3.400E+0
VARIANCE    =  5.605E+01    SKEWNESS=  6.620E-01    KURTOSIS=  3.053E+0
COEFF. VAR. =  4.737E+01    AVG.DEV.=  5.895E+00    RMS DEV.=  7.382E+0



          6 CELLS - CELL INTERVAL = 6

 MIDPOINT    NO. OBS.      % TOTAL      TOT.CUM.     Z-SCORE(RMS)
 6.000          8          21.622       21.622         -1.328
 1.200E+01     11          31.081       52.703          -.516
 1.800E+01      9          25.676       78.378           .297
 2.400E+01      3           9.459       87.838          1.110
 3.000E+01      2           5.405       93.243          1.923
 3.600E+01      1           4.054       97.297          2.736
```


5
10
15
20
25
30
35
40
99.99
60
50
40
30
20
10
5
Arith. Histogram


LOG - PROBABILITY
K&E PROBABILITY x 2 LOG CYCLES 46 8040
99.99
99.9
99.8
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
0.01
10
9
8
7
6
5
4
3
2
1
9
8
7
6
5
4
3
2

```
SAMPLE STATISTICS

TEST FRIEDMAN DATA   PROBE 2

OBSERVATIONS=  3.600E+01    MEAN      =  1.178E+01     STD.DEV.=  4.934E+00
MINIMUM     =  5.000E+00    RANGE     =  2.000E+01     MAXIMUM =  2.500E+01
VARIANCE    =  2.435E+01    SKEWNESS=  8.326E-01      KURTOSIS=  3.545E+00
COEFF. VAR. =  4.190E+01    AVG.DEV.=  3.778E+00      RMS DEV.=  4.865E+00


#

INTERRUPT DURING LINE 1680
>RUN
  DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION?%FRIED
  JOB TITLE - NO. ROWS -NO. COLUMNS?TEST FRIEDMAN DATA--LOG TRANSFORM,36,10
  LOG TRANSFORM?  INPUT  1 FOR YES  2 FOR NO ?1
```

NONPARAMETRIC SUMMARY PROGRAM

There is a tremendous effort in terms of manpower and costs involved in designing and collecting air quality data. All of this information should be summarized in air quality reports. The following output is recommended for all air quality reports.

	SITES					
HOUR	1	2	3	4	5	6
0600	A	A	A	A	A	A
	B	B	B	B	B	B
	C	C	C	C	C	C
0700	A	A	A	A	A	A
	B	B	B	B	B	B
	C	C	C	C	C	C
0800	A	A	A	A	A	A
	B	B	B	B	B	B
	C	C	C	C	C	C
0900	A	A	A	A	A	A
	B	B	B	B	B	B
	C	C	C	C	C	C

A = minimum or lower confidence limits ($\alpha = .05$).

B = median (not average).

C = maximum or upper confidence limit ($\alpha = .05$).

This type of summary table should be included in all air quality reports for each month of sampling.

You should explain why the minimum and maximum value occurred either through traffic or meteorology.

It is recommended that all districts use this summary with minimum and maximum value. However, for air quality studies in large metropolitian areas it may be desirable to use the 95% confidence limits rather than maximum and maximum values. This format should be used for CO, CH, NO_x, and Ozone.

For those districts not using DIFKIN, a trend analysis based on the pollutant burden concept can be used to indicate whether these values will increase or decrease in future years.

Example:

```
>LINK '5;LSTAT;SUMMARY'
     ***** MODIFIED SEPTEMBER, 1974 *****

     DATA FILENAME OR 'EXP' FOR PROGRAM DETAILS?*FRIED

     NUMBER OF APCD'S, NUMBER OF SITES?2,8
     NUMBER OF DAYS?6
     NUMBER OF READINGS/DAY AT EACH SITE?6
     EARLIEST READING (MILITARY TIME)?0700

     SITE ID, MINUTES AFTER HOUR READ (2 CHAR EACH)
       APCD 1?01,00
       APCD 2?02,00
       SITE 1?01,00
       SITE 2?02,00
       SITE 3?03,05
       SITE 4?04,05
       SITE 5?05,00
       SITE 6?06,00
       SITE 7?07,00
       SITE 8?08,00

    INPUT 6 LINES FOR HEADING (0=BLANK LINE)
    LINE 1?0
    LINE 2?SAMPLE RUN OF THE
    LINE 3?0
    LINE 4?NON-PARAMETRIC SUMMARY PROGRAM
    LINE 5?0
    LINE 6?0

    0=MIN & MAX, 1=LIMITS - WHICH?1
    SIGNIFICANCE LEVEL?.05
```

SAMPLE RUN OF THE

NON-PARAMETRIC SUMMARY PROGRAM

```
HOUR      APCD  APCD  SITE  SITE  SITE  SITE  SITE  SITE  SITE  SITE
           01    02    01    02    03    04    05    06    07    08
          H+00  H+00  H+00  H+00  H+05  H+05  H+00  H+00  H+00  H+00
0700
     N     6     6     6     6     6     6     6     6     6     6
     UL   21.0  18.0  10.0  14.0  21.0  14.0  14.0  16.0  11.0  19.0
     MED  10.5   8.0   3.5   6.5   6.0   8.0  10.5   9.0   6.5  10.5
     LL    5.0   5.0   3.0   4.0   4.0   3.0   6.0   4.0   5.0   8.0
0800
     N     6     6     6     6     6     6     6     6     6     6
     UL   33.0  25.0  14.0  22.0  25.0  20.0  16.0  22.0  15.0  18.0
     MED  18.5  13.0   5.0  10.5   9.0  10.0  14.0  14.0  10.5  12.5
     LL    4.0   5.0   3.0   5.0   5.0   4.0   5.0   4.0   8.0  11.0
0900
     N     6     6     6     6     6     6     6     6     6     6
     UL   23.0  20.0  14.0  17.0  22.0  17.0  19.0  20.0  14.0  17.0
     MED  13.0  12.0   7.0   8.0   9.0   7.5   9.5  13.0   9.5   9.5
     LL    6.0   5.0   2.0   5.0   5.0   3.0   5.0   6.0   8.0   8.0
1000
     N     6     6     6     6     6     6     6     6     6     6
     UL   30.0  13.0  15.0  22.0  24.0  17.0  24.0  19.0  16.0  24.0
     MED  16.5  10.0   6.5  11.0  10.0   9.0  12.5  10.0   8.0  13.5
     LL    7.0   6.0   2.0   4.0   5.0   4.0   8.0   8.0   5.0   8.0
1100
     N     6     6     6     6     6     6     6     6     6     6
     UL   34.0  16.0  24.0  29.0  27.0  30.0  30.0  23.0  21.0  21.0
     MED  19.0  13.5  10.5  11.5  14.0  14.0  16.0  14.0  10.5  18.0
     LL   13.0   8.0   3.0   4.0   6.0   5.0   9.0  14.0   8.0   9.0
1200
     N     6     6     6     6     6     6     6     6     6     6
     UL   27.0  24.0  22.0  25.0  28.0  22.0  27.0  22.0  17.0  18.0
     MED  14.0  10.0   9.0  10.5  10.5   9.5  11.0  13.5  11.5  13.0
     LL    7.0   8.0   3.0   3.0   5.0   4.0   6.0   8.0   9.0   8.0
1300
```

COMPUTER PROGRAM
FOR LARSEN'S MODEL

The following pages contain example computer runs of Larsen's Mathematical Model for relating air quality measurements to air quality standards. The model can be used for calculating the following pollutant parameters, for any averaging time: geometric mean, standard geometric deviation, maximum concentration expected once a year, and frequency distribution of expected pollutant concentrations. Future year concentrations are projected by proportional modeling as described earlier in this manual.

Access to the Tenet version is gained by the basic command:

```
LINK "5;LAB;LARSEN"
```

A complete explanation of the operating procedures for this program can be attained by answering "yes" when the program asks if the user would like a program explanation.

The present version of the Larsen Model will accept input data in three forms:

1. A pollutant data file.
2. Previously generated pollutant statistics.
3. A pollutant frequency distribution.

The first example demonstrates the form of input and output for a pollutant data file. The file used in this example is named "CODATA". The file is shown on the following page. The actual computer run follows on page 245. The output format is designed so that the user can use the output as generated in an air quality report, if desired.

A plot of the observed CO distribution is shown on page 248. The Larsen graphical solution is shown on the same plot so that a comparison can be made. Blank forms are provided on pages 261 and 262 for sample plots.

```
                              02/04/75

5;LAB;LARSEN--FOR DATA STATISTICS AND/OR LARSEN'S MODEL
LATEST REVISION - 1/31/75

QUESTIONS? CALL PAUL ALLEN 8-432-4877

PROGRAM EXPLANATION (YES OR NO) ?NO


JOB TITLE = ?CO DATA FILE TEST RUN

ARE YOU INPUTTING A DATA FILE (YES OR NO) ?YES

FILE NAME = ?CODATA

NUMBER OF ROWS IN YOUR FILE =?34

NUMBER OF COLUMNS IN YOUR FILE =?10


HOW MANY COLUMNS TO BE INCLUDED IN THIS RUN ?10


INTERVAL FOR DATA SEARCH (CELL INTERVAL) = ?1

DO YOU WISH DATA LISTING (YES OR NO) ?YES
```

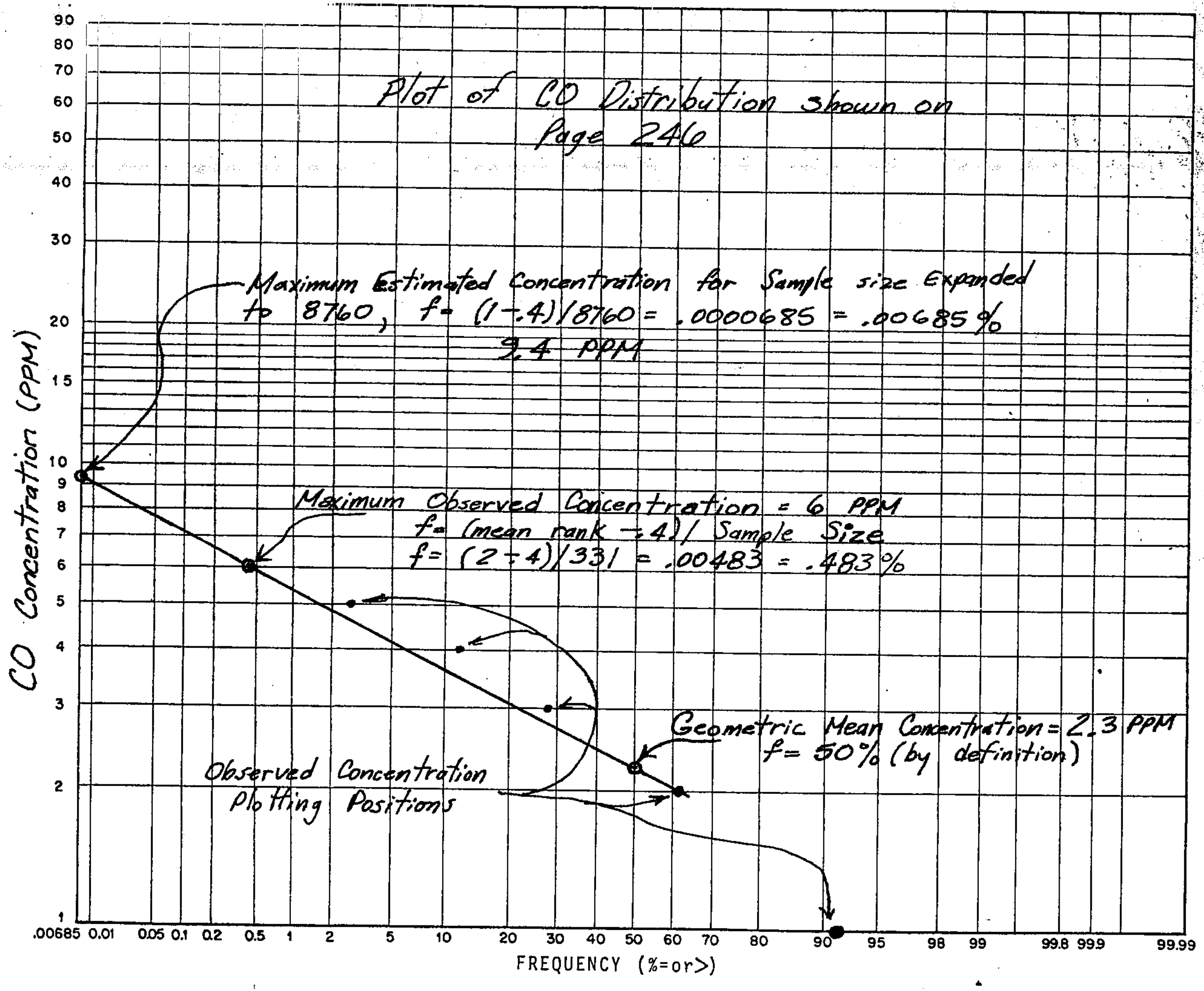
02/04/75

CO DATA FILE TEST RUN

```
DATA

         1      2      3      4      5      6      7      8      9     10

***************************************************************************

  1--   6.00   6.00   6.00   5.00   5.00   5.00   5.00   5.00   5.00   5.00
  2--   5.00   5.00   5.00   5.00   5.00   5.00   5.00   5.00   4.00   4.00
  3--   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
  4--   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
  5--   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
  6--   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00   4.00
  7--   4.00   4.00   4.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
  8--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
  9--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
 10--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
 11--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
 12--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00
 13--   3.00   3.00   3.00   3.00   3.00   3.00   3.00   3.00   2.00   2.00
 14--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 15--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 16--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 17--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 18--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 19--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 20--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 21--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 22--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 23--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 24--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 25--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 26--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 27--   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00   2.00
 28--   2.00   2.00   2.00   2.00   2.00   2.00   1.00   1.00   1.00   1.00
 29--   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
 30--   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
 31--   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
 32--   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
 33--   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
 34--   1.00  -1.00  -1.00  -1.00  -1.00  -1.00  -1.00  -1.00  -1.00  -1.00
```

```
                              02/04/75

                         CO DATA FILE TEST RUN

ARITHMETIC MEAN =   2.474
STANDARD ARITHMETIC DEVIATION =   1.113

RANGE IS     1.00  TO     6.00
MEDIAN =    2.00

SAMPLE SIZE = 331


                           FREQUENCY DISTRIBUTION

           FREQUENCY OF      CUMULATIVE       LARSEN'S       NUMBER OF
            OCCURRENCE       FREQUENCY        FREQUENCY      OCCURRENCES
               (%)            % =OR<           % =OR>

 1.00--      16.616           16.616           91.722           55.0
 2.00--      44.713           61.329           61.057          148.0
 3.00--      19.637           80.967           28.882           65.0
 4.00--      13.595           94.562           12.266           45.0
 5.00--       4.532           99.094            3.202           15.0
 6.00--        .906          100.000             .483            3.0


FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN: (*)

    STANDARD GEOMETRIC DEVIATION =  1.446
    GEOMETRIC MEAN =  2.312

FROM ALL OBSERVED DATA:

    STANDARD GEOMETRIC DEVIATION =  1.536
    GEOMETRIC MEAN =  2.257

(*) REFERENCE AP-89
```

```
USE LARSEN'S FREQUENCY FOR PLOTTING ON LOG-PROBABILITY PAPER.


DO YOU WANT A LARSEN'S MODEL ANALYSIS
ON YOUR DATAFILE AT THIS TIME(YES OR NO)?YES



DO YOU WANT A FREQUENCY LISTED FOR EACH VALUE IN THE
LARSEN DISTRIBUTION (YES OR NO) ?YES

DO YOU WANT AN ANALYSIS FOR 1-HR AVE. TIME ?YES

INPUT STANDARD FOR 1-HR AVE. TIME (SAME UNITS AS DATA)?35

WHAT UNITS ARE YOUR DATA IN(PPM OR PPHM)?PPM

HOW MANY OTHER AVERAGING TIMES DO YOU WANT ANALYZED ?1

INPUT THE AVERAGING TIMES (HOURS) ?8

INPUT THE RESPECTIVE STANDARDS(SAME UNITS AS DATA)?9


WHAT SAMPLE SIZE FOR 1-HR AVE. TIME DO YOU WANT
YOUR SAMPLE EXPANDED TO (HOURS) ?8760
```



Plot of CO Distribution Shown on Page 246

```
                    02/04/75

               CO DATA FILE TEST RUN


  STATISTICAL PARAMETERS FROM 1-HOUR AVERAGING TIME DATA

           ARITHMETIC MEAN =  2.474
           STANDARD DEVIATION =  1.113


       FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN:

           STANDARD GEOMETRIC DEVIATION =  1.446
           GEOMETRIC MEAN =  2.312


       EXPECTED MAXIMUM 1-HOUR CONCENTRATION
       FROM SAMPLING PARAMETERS =  9.43 PPM


       NUMBER OF TIMES AMBIENT AIR QUALITY
       STANDARD OF  35.00 PPM  TO BE EXCEEDED =     0 PER YEAR


   PREDICTED LARSENS FREQUENCIES FOR 1-HOUR CONCENTRATIONS

        CONCENTRATION                  LARSENS FREQUENCY
            PPM                             % =OR>

            1.00                           98.8521
            2.00                           65.2925
            3.00                           23.9778
            4.00                            6.8421
            5.00                            1.8167
            6.00                             .4828
            7.00                             .1323
            8.00                             .0378
            9.00                             .0113

  SAMPLE SIZE = 8760

  ALL CALCULATIONS REFERENCE AP-89
```

```
                              02/04/75

                     CO DATA FILE TEST RUN

 STATISTICAL PARAMETERS FROM  8-HOUR AVERAGING TIME DATA

              GEOMETRIC MEAN =  2.325

              STANDARD GEOMETRIC DEVIATION =  1.382

          EXPECTED MAXIMUM     8-HOUR CONCENTRATION
          FROM SAMPLING PARAMETERS =  6.76 PPM


          NUMBER OF TIMES AMBIENT AIR QUALITY
          STANDARD OF  9.00 PPM  TO BE EXCEEDED =    0 PER YEAR

   PREDICTED LARSENS FREQUENCIES FOR    8-HOUR CONCENTRATIONS

           CONCENTRATION
               PPM                       LARSENS FREQUENCY
                                              % =OR>
              1.00
              2.00                            99.5443
              3.00                            67.9239
              4.00                            21.5490
              5.00                             4.6811
              6.00                              .8985
              7.00                              .1697
                                                .0330

SAMPLE SIZE = 1095

ALL CALCULATIONS REFERENCE AP-89
```

The next run uses the statistics generated from the data file to demonstrate the form of input for an analysis on statistics generated previous to the run. Generally, this will not be needed since it is anticipated that pollutant monitoring data will usually be used. In any case the distribution should be plotted on log-probability paper to check the applicability of the Larsen approach.

```
                    02/04/75

5;LAB;LARSEN--FOR DATA STATISTICS AND/OR LARSEN'S MODEL
LATEST REVISION - 1/31/75

QUESTIONS? CALL PAUL ALLEN 6-432-4877

PROGRAM EXPLANATION (YES OR NO) ?NO


JOB TITLE = ?CO DATA TEST USING FRQUENCY DISTRIBUTION INPUT

ARE YOU INPUTTING A DATA FILE (YES OR NO) ?NO

DO YOU HAVE DISTRIBUTION STATS(1) OR A FREQ DIST.(2)?2



INPUT THE NUMBER OF INTERVALS IN THE DISTRIBUTION?6

ENTER THE CONCENTRATION AND ITS FREQUENCY (LO TO HI)
(ENTER FREQUENCIES IN PERCENTAGES; I.E. TYPE 35 FOR
35% - SUM OF ALL FREQS SHOULD APPROXIMATE 100.)

FOR INTE1,16.616
L NO.    1
FOR INTERVAL NO.     2?2,44.713

FOR INTERVAL NO.     3?5,19.637

FOR INTERVAL NO.     4?4,13.595

FOR INTERVAL NO.     5?5,4.532

FOR INTERVAL NO.     6?6,.906

INPUT THE SAMPLE SIZE FOR THE DISTRIBUTION?331
INTERVAL FOR DATA SEARCH (CELL INTERVAL) = ?1
```

```
                         02/04/75

          CO DATA TEST USING FRQUENCY DISTRIBUTION INPUT

ARITHMETIC MEAN =   2.474
STANDARD ARITHMETIC DEVIATION =   1.113

RANGE IS     1.00  TO     6.00
MEDIAN =    2.00

SAMPLE SIZE = 331


                         FREQUENCY DISTRIBUTION

            FREQUENCY OF    CUMULATIVE     LARSEN'S       NUMBER OF
             OCCURRENCE     FREQUENCY      FREQUENCY     OCCURRENCES
                (%)          % =OR<         % =OR>

 1.00--        16.616         16.616         91.722           55.0
 2.00--        44.713         61.329         61.057          148.0
 3.00--        19.637         80.966         28.882           65.0
 4.00--        13.595         94.561         12.266           45.0
 5.00--         4.532         99.093          3.202           15.0
 6.00--          .906         99.999           .483            3.0



FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN: (*)

    STANDARD GEOMETRIC DEVIATION =  1.446
    GEOMETRIC MEAN =  2.312

FROM ALL OBSERVED DATA:

    STANDARD GEOMETRIC DEVIATION =  1.536
    GEOMETRIC MEAN =  2.257

(*) REFERENCE AP-89
```

```
USE LARSEN'S FREQUENCY FOR PLOTTING ON LOG-PROBABILITY PAPER.


DO YOU WANT A LARSEN'S MODEL ANALYSIS
ON YOUR DATAFILE AT THIS TIME(YES OR NO)?YES


DO YOU WANT A FREQUENCY LISTED FOR EACH VALUE IN THE
LARSEN DISTRIBUTION (YES OR NO) ?NO


DO YOU WANT AN ANALYSIS FOR 1-HR AVE. TIME ?YES

INPUT STANDARD FOR 1-HR AVE. TIME (SAME UNITS AS DATA)?35


HOW MANY OTHER AVERAGING TIMES DO YOU WANT ANALYZED ?1

INPUT THE AVERAGING TIMES (HOURS) ?8

INPUT THE RESPECTIVE STANDARDS(SAME UNITS AS DATA)?9


WHAT SAMPLE SIZE FOR 1-HR AVE. TIME DO YOU WANT
YOUR SAMPLE EXPANDED TO (HOURS) ?8760
WHAT UNITS ARE YOUR DATA IN(PPM OR PPHM)?PPM
```

```
                         02/04/75

           CO DATA TEST USING FRQUENCY DISTRIBUTION INPUT


    STATISTICAL PARAMETERS FROM 1-HOUR AVERAGING TIME DATA

              ARITHMETIC MEAN =  2.474
              STANDARD DEVIATION =  1.113


         FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN:

              STANDARD GEOMETRIC DEVIATION =  1.446
              GEOMETRIC MEAN =  2.312


         EXPECTED MAXIMUM 1-HOUR CONCENTRATION
         FROM SAMPLING PARAMETERS =  9.43 PPM


         NUMBER OF TIMES AMBIENT AIR QUALITY
         STANDARD OF  35.00 PPM  TO BE EXCEEDED =    0 PER YEAR

   SAMPLE SIZE = 8760

   ALL CALCULATIONS REFERENCE AP-89
```

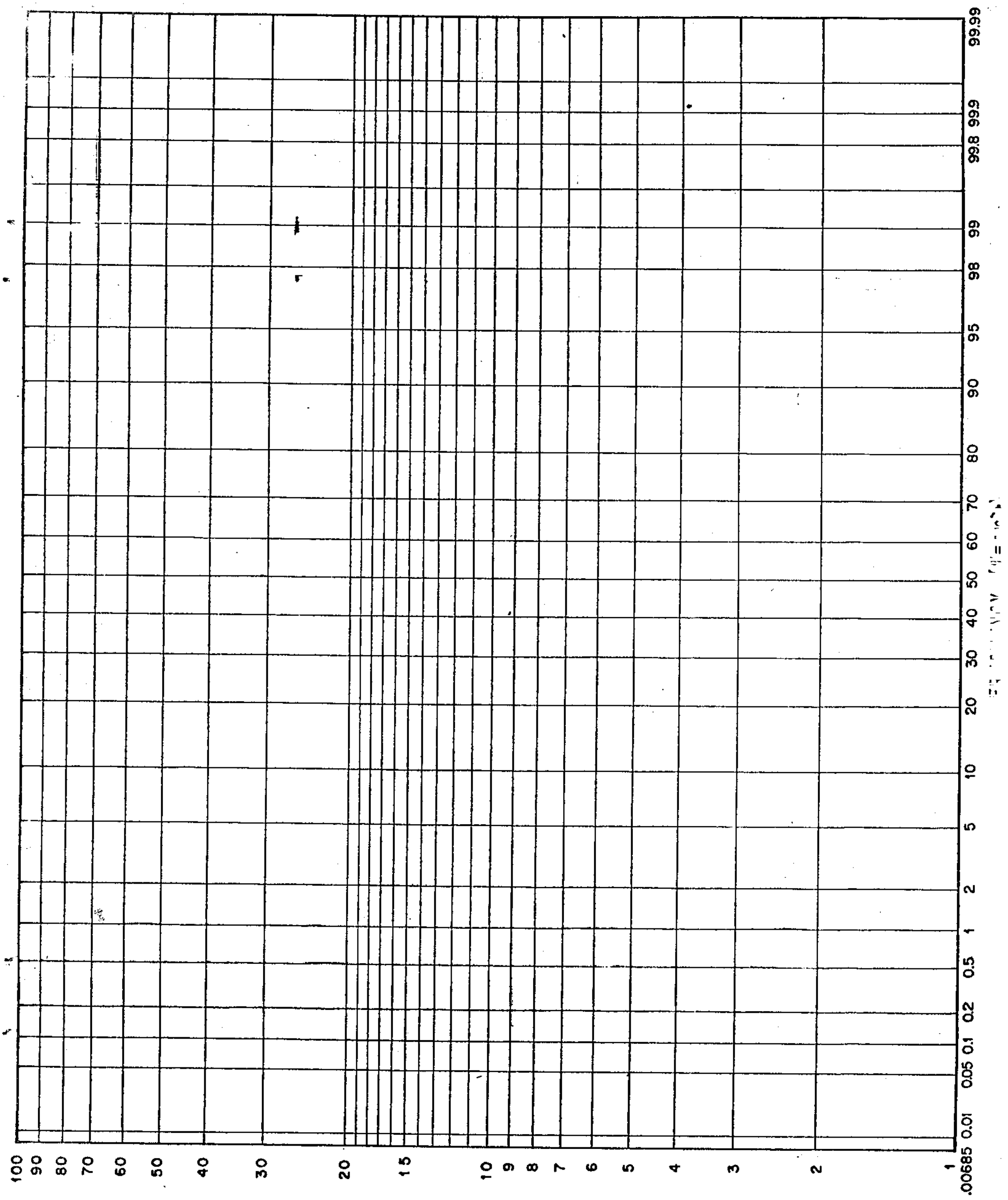
```
                    02/04/75

     CO DATA TEST USING FRQUENCY DISTRIBUTION INPUT


STATISTICAL PARAMETERS FROM  8-HOUR AVERAGING TIME DATA

          GEOMETRIC MEAN =  2.325

          STANDARD GEOMETRIC DEVIATION =  1.382


      EXPECTED MAXIMUM    8-HOUR CONCENTRATION
      FROM SAMPLING PARAMETERS =  6.76 PPM



      NUMBER OF TIMES AMBIENT AIR QUALITY
      STANDARD OF  9.00 PPM  TO BE EXCEEDED =    0 PER YEAR

SAMPLE SIZE = 1095

ALL CALCULATIONS REFERENCE AP-89
```

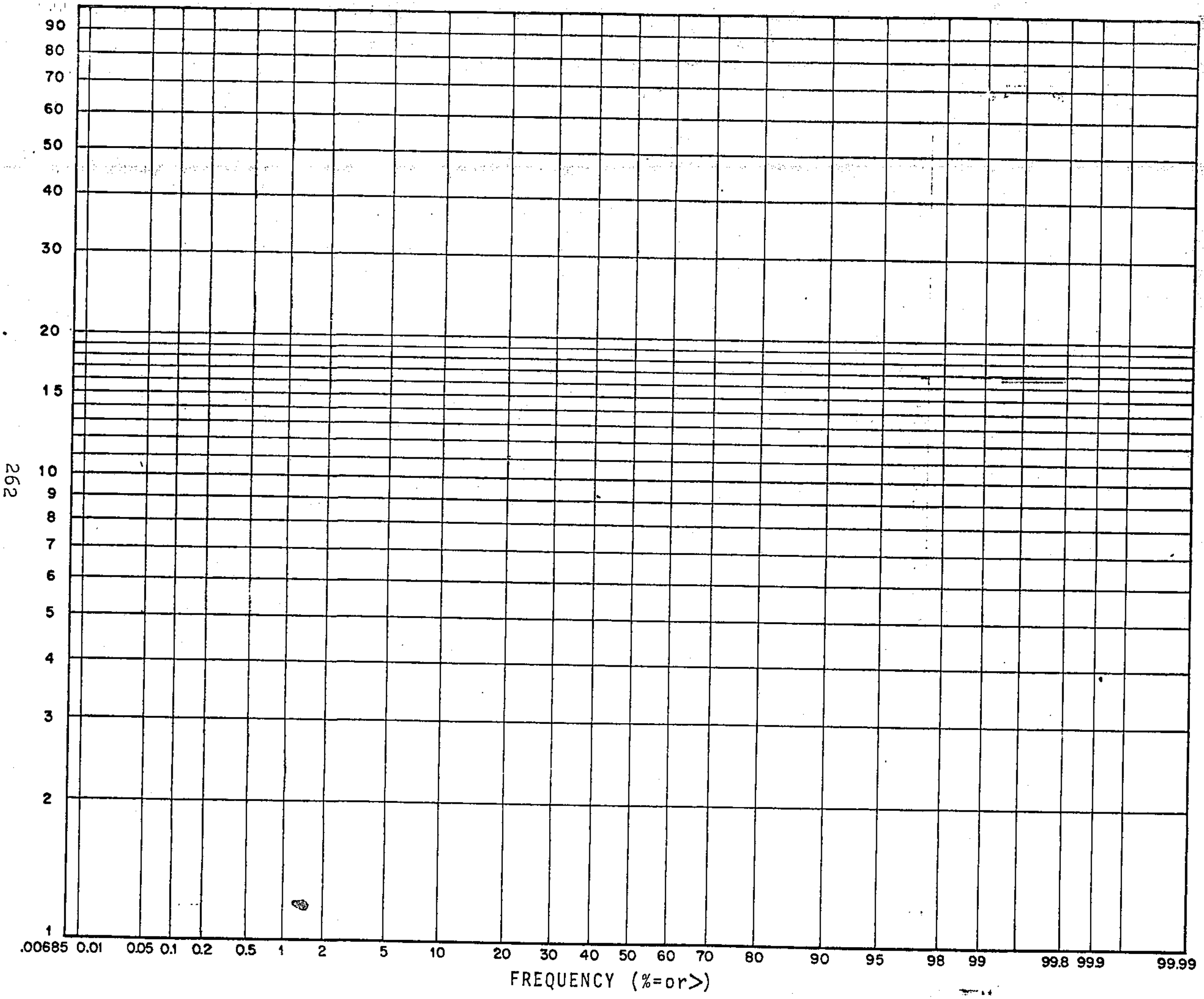
The final run also uses the statistics generated from the first run. The frequency distribution is input. This method lends itself to a Larsen analysis on a frequency distribution generated according to Part VII of this manual.

```
02/04/75

5;LAB;LARSEN--FOR DATA STATISTICS AND/OR LARSEN'S MODEL
LATEST REVISION - 1/31/75

QUESTIONS? CALL PAUL ALLEN 8-432-4877

PROGRAM EXPLANATION (YES OR NO) ?NO


JOB TITLE = ?CO STAT TEST RUN

ARE YOU INPUTTING A DATA FILE (YES OR NO) ?NO

DO YOU HAVE DISTRIBUTION STATS(1) OR A FREQ DIST.(2)?1

INPUT MAXIMUM OBSERVED CONCENTRATION?6

INPUT NUMBER OF OCCURRENCES OF THIS MAX.?3

INPUT THE SAMPLE SIZE?331

INPUT THE ARITHMETIC MEAN?2.474

INTERVAL FOR DATA SEARCH (CELL INTERVAL) = ?1

DO YOU WANT A FREQUENCY LISTED FOR EACH VALUE IN THE
LARSEN DISTRIBUTION (YES OR NO) ?YES

DO YOU WANT AN ANALYSIS FOR 1-HR AVE. TIME ?YES

INPUT STANDARD FOR 1-HR AVE. TIME (SAME UNITS AS DATA)?35


HOW MANY OTHER AVERAGING TIMES DO YOU WANT ANALYZED ?1

INPUT THE AVERAGING TIMES (HOURS) ?8

INPUT THE RESPECTIVE STANDARDS(SAME UNITS AS DATA)?9


WHAT SAMPLE SIZE FOR 1-HR AVE. TIME DO YOU WANT
YOUR SAMPLE EXPANDED TO (HOURS) ?8760
WHAT UNITS ARE YOUR DATA IN(PPM OR PPHM)?PPM
```

```
                          02/04/75

                       CO STAT TEST RUN


    STATISTICAL PARAMETERS FROM 1-HOUR AVERAGING TIME DATA

            ARITHMETIC MEAN =  2.474
            STANDARD DEVIATION =    .944


        FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN:

            STANDARD GEOMETRIC DEVIATION =  1.446
            GEOMETRIC MEAN =  2.312


        EXPECTED MAXIMUM 1-HOUR CONCENTRATION
        FROM SAMPLING PARAMETERS =  9.43 PPM


        NUMBER OF TIMES AMBIENT AIR QUALITY
        STANDARD OF  35.00 PPM  TO BE EXCEEDED =    0 PER YEAR


    PREDICTED LARSENS FREQUENCIES FOR 1-HOUR CONCENTRATIONS

         CONCENTRATION                 LARSENS FREQUENCY
              PPM                           % =OR>

              1.00                          98.8497
              2.00                          65.2751
              3.00                          23.9686
              4.00                           6.8398
              5.00                           1.8164
              6.00                            .4828
              7.00                            .1323
              8.00                            .0378
              9.00                            .0113

   SAMPLE SIZE = 8760

   ALL CALCULATIONS REFERENCE AP-89
```

```
                         02/04/75

                     CO STAT TEST RUN


STATISTICAL PARAMETERS FROM  8-HOUR AVERAGING TIME DATA

          GEOMETRIC MEAN =  2.325

          STANDARD GEOMETRIC DEVIATION =  1.382


      EXPECTED MAXIMUM    8-HOUR CONCENTRATION
      FROM SAMPLING PARAMETERS =  6.76 PPM



      NUMBER OF TIMES AMBIENT AIR QUALITY
      STANDARD OF  9.00 PPM  TO BE EXCEEDED =    0 PER YEAR


  PREDICTED LARSENS FREQUENCIES FOR    8-HOUR CONCENTRATIONS


       CONCENTRATION                  LARSENS FREQUENCY
          PPM                              % =OR>

         1.00                              99.5431
         2.00                              67.9049
         3.00                              21.5393
         4.00                               4.6792
         5.00                                .8983
         6.00                                .1697
         7.00                                .0330


SAMPLE SIZE = 1095


ALL CALCULATIONS REFERENCE AP-89
```


100
90
80
70
60
50
40
30
20
15
10
9
8
7
6
5
4
3
2
1
.00685
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99


90
80
70
60
50
40
30
20
15
10
9
8
7
6
5
4
3
2
1
.00685
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
FREQUENCY (%=or>)

HOMEWORK ASSIGNMENTS

HOMEWORK PROBLEM NO. 1

The following frequency table summarized a sample of 286 hourly CO measurements at a site of a specified study area over a period of one season:

CO (ppm), Xi	1	2	3	4	5	6	7	8	9	10	11
Frequency, fi	20	40	68	50	40	25	20	10	8	4	1

1. Find, using equations:

(a)	The median concentration	Answers: (a)	4.25
(b)	The mean concentration, m	(b)	4.19
(c)	The geometric mean concentration, mg	(c)	3.64
(d)	The standard deviation, s	(d)	2.11
(e)	The standard geometric deviation, Sg	(e)	1.74

2. Test (by graphical method) whether or not the above set of CO concentrations follows approximately a lognormal distribution.

3. If the distribution is lognormal, find Mg and Sg from the lognormal probability paper and compare these estimates with the values calculated in 1-(c) and 1-(e) above.

Hint:

X	1	2	3	4	5	6	7	8	9	10	11	1.74	3.64
ln X		0.693	1.099	1.386	1.609	1.792	1.946	2.080	2.197	2.303	2.398	.554	1.29

HOMEWORK PROBLEM NO. 2

The air quality data for an APCD station for the month of July 1971 measured 15 days where the O_3 concentrations exceeded the Federal Health Standard of 0.08 ppm. In the following year (July 1972) the station observed 7 days where O_3 concentrations exceeded the federal standard.

The APCD station also has a wind system which measures surface wind speeds and directions.

Hourly wind roses were analyzed for wind speed and direction for the time period from 1400 to 1800 for both years for the month of July.

	Site 1	Site 2
Time	$\bar{U}$ (mph)	$\bar{U}$ (mph)
1400-1500	8	10
1500-1600	12	11
1600-1700	11	9
1700-1800	10	8

Answer the following:

1. What important parameters affect the 0_3 concentrations?

2. What may be the possible cause(s) of the reduction of 0_3? Explain.

3. Which test(s) can be used to analyze these data to determine if there is a significant difference in wind speed?

HOMEWORK PROBLEM NO. 3

A highway route is proposed near an airport. Based on 20 years of historical meteorological data the most probable surface stability for a summer condition was Stability Class D with a prevailing wind direction from the East. These conditions were estimated for the peak morning traffic (0700-0900) hours. The most frequent wind speed interval for the easterly direction was 4 to 7 mph. These meteorological conditions are to be used as inputs into the line source dispersion model to estimate the CO concentrations above background. An APCD station is located approximately 2 miles from the proposed route. The exposure of the air monitoring station was considered marginal; therefore, it was decided that an ambient air quality survey should be made for CO.

There were 8 different sampling locations for CO. All sampling locations were located about 2 miles apart. All sampling at each site began at the same time. All samples were obtained for a one-hour averaging time. Sampling began at 0700 and ended at 1800.

The following is a summary of the CO concentration measured at the sites when the most probable summer meteorological conditions of stability, wind speed and direction were met.

CO CONCENTRATIONS IN PPM

			Sampling Sites							
Date	Time	APCD	1	2	3	4	5	6	7	8
8/15/72	0700	12	8	9	*	9	8	12	9	8
	0800	5	5	5	6	5	5	*	6	6
8/16/72	0700	7	5	5	4	4	5	10	8	8
	0800	4	3	4	3	4	5	5	4	7
8/17/72	0700	8	4	6	4	4	5	8	7	7
	0800	6	4	4	3	5	4	4	4	4
8/18/72	0700	3	3	4	3	3	3	3	*	3
	0800	3	3	3	3	3	3	3	3	3
8/19/72	0700	2	2	3	2	2	3	*	2	2
	0800	2	2	3	2	2	3	3	2	2
8/23/72	0700	11	7	9	6	8	*	12	9	14
	0800	11	8	10	6	4	6	7	7	5
8/30/72	0700	4	4	5	3	3	4	3	3	4
	0800	4	3	*	3	4	3	3	3	3

*No data, air sampling bag leaked.

All measurements made between 0700 to 0900.

Sampling error ± 0.5 ppm.

Answer the following:

1. What are the important parameters affecting the CO concentrations at each site?

2. Which test(s) can be used to determine if there is no significance difference in the spatial distribution of CO measured at all sites (including the APCD station).

3. If the APCD station is not representative of all the sites, how would you determine if it is representative of any individual site or sites?

HOMEWORK PROBLEM NO. 4

An ambient air quality survey was planned for CO. Sampling began on September 1, 1972. Three sites were selected based on homogeneous areas and the general criteria described in the ambient air quality manual. All air samples were taken simultaneously at all three sites to determine the temporal and spatial distribution of CO. Sampling began at 0600 and ended at 1900. Sampling for the project was based on a balance randomized block design. A mechanical weather station was installed to measure wind speeds and directions. The data from this station were not available until the end of the sampling period.

The following is a summary of the first two days of CO measurements for the morning period (0600-0900).

September 1, 1972

Time	Site 1	Site 2	Site 3
0600-0700	6	7	4
0700-0800	5	8	4
0800-0900	3	6	2

All values are in ppm

September 6, 1972

Time	Site 1	Site 2	Site 3
0600-0700	8	5	5
0700-0800	8	6	8
0800-0900	4	7	3

Answer the following:

1. What parameters influence the background level of CO at each site?

2. Which test(s) can be used to analyze these data to determine whether there is a significant difference in the spatial distribution of CO measured at all sites for these two days sampled for the hours of 0600 to 0900?

3. If there is no significant difference in the CO concentrations measured on each of these days does this infer that the samples were taken under similar meteorological conditions? Explain.

4. Could a statistical analysis be made for each of the days for the entire sampling period from 0600 to 1900? Explain

HOMEWORK PROBLEM NO. 5

Before beginning an ambient air quality survey for CO, it was decided by District 07 to correlate the Transportation Department's test procedure of bag sampling to the nearest APCD station which monitors continuously. In this study air samples obtained by District 07 and the APCD station were taken at the same location to assure that both were sampling the same air. The following is a summary of the data collected and analyzed by the Department of Transportation and the APCD station, respectively:

Time	Division of Highways	APCD Station
0800-0900	6	4
0900-1000	5	4
1000-1100	4	4
1100-1200	4	3
1200-1300	4	3
1300-1400	4	3
1400-1500	3	3

All values in ppm.

Answer the following:

1. What important parameters should be considered for this correlation?

2. Which test(s) can be used to analyze these data to determine if there is a significant difference in sampling procedures?

3. Is it necessary to consider the changes in meteorology and traffic volumes in analyzing these data? Explain.

HOMEWORK PROBLEM NO. 6

1. What is the equation for the standard geometric deviation using all the data?

2. What is the equation for the geometric mean using all the data?

3. What percentile is the once-a-year one-hour standard related to? (hint---there are 8760 hours in a year and AP-89 has Table 11)

4. What type of distribution is used for aerometric pollutants and why?

5. In Larsen's Model what physical relationship do the following variables have: Arithmetic mean, m?; Standard geometric deviation, Sg?

6. What is meant by the worst/worst hour air quality impact for transportation air quality impact reports on the microscale level?

7. Larsen's Model is good for what type of pollutants?

8. Why is Larsen's Model suggested in the reporting of air quality impacts transportation projects?

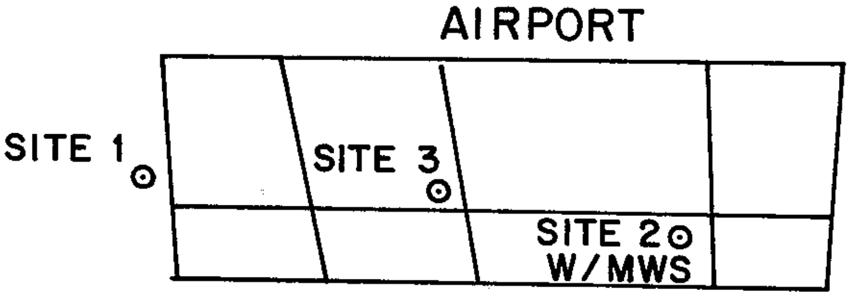
9. When should Larsen's Model not be used?

10. What should you always do when using Larsen's Model?

11. What part of the collected data is the most important when analyzing the field data for use in Larsen's Model?

12. You have the arithmetic mean of a long-term series of CO concentration measurements---approximately what position would this be on lognormal probability paper in percentile if the measured concentrations are of one-hour measurements?

HOMEWORK PROBLEM NO. 7



We have a mobile air monitoring laboratory which must sample the three sites shown in the diagram. These sites are all located within 20 miles of one another, and all must make use of the mobile laboratory.

1. Organize a sampling scheme for oxidant (covering the month of August) using a randomized block design.

2. List of steps that you would undertake to generate the missing data (days not sampling at a given site).

HOMEWORK PROBLEM NO. 8

The following pages contain a trial run of STPREG. The data that was used is the same as the earlier STPREG run except that logs were taken of all the pollutant data. Consider the importance of the statistical parameters. Compare this trial to the earlier run. Which data arrangement gives the better statistical prediction equation?

```
STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE APCD [redacted] DATA TO VAN DATA--LOG-LOG TRANS. TEST

NUMBER OF OBSERVATIONS     42
NUMBER OF VARIABLES         6
FORCE ZERO INTERCEPT       NO



VARIABLE                  MEAN          VARIANCE          STD. DEV.

 TIME/100   1           12.50000        16.64634          4.07999
FIELD OZ.   2           -1.48370          .04760           .21818
 APCD OZ.   3           -1.53481          .04532           .21289
 SKY CODE   4            2.38095         1.60743          1.26785
  INV/100   5           19.85714        65.10105          8.06852
     UBAR   6            4.28571         2.89199          1.70058


CORRELATION MATRIX

ROW 1
   1.00000

ROW 2
    .08029   1.00000

ROW 3
    .06977    .82658   1.00000

ROW 4
    .10845    .28774    .10490   1.00000

ROW 5
    .87056    .25936    .18430    .33686   1.00000

ROW 6
    .25661    .83558    .65653    .35553    .39411   1.00000
 CONTROL DATA FOR PROBLEM NO.  1

 F-LEVEL FOR INCLUSION (0=.01)?0
 F-LEVEL FOR DELETION (0=.005)?0
 TOLERANCE LEVEL (0=.001)?0

 VARIABLE CONROL VALUES
    0 - DELETE VARIABLE FROM ANALYSIS
    1 - DEPENDENT VARIABLE
    2 - FREE VARIABLE - MAY BE USED IN ANALYSIS
    3 TO 9 - FORCED VARIABLE - LOW TO HIGH LEVEL

 CONTROL VALUE FOR:
 TIME/100  1 = ?2
FIELD OZ.  2 = ?1
 APCD OZ.  3 = ?2
 SKY CODE  4 = ?2
  INV/100  5 = ?2
     UBAR  6 = ?2

 THIS PROBLEM MAY REQUIRE UP TO 10 STEPS TO SOLVE.

 ENTER THE MAXIMUM NUMBER OF STEPS DESIRED FOR SOLUTION?10
```

STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE ~~FIELD~~ APCD DATA TO VAN DATA--LOG-LOG TRANS. TEST

```
PROBLEM NUMBER            1                      F TO ENTER           .01000
STEP NUMBER               1                      F TO REMOVE          .00500
DEPENDENT VARIABLE        2 (FIELD OZ.)          TOLERANCE LEVEL      .00100

VARIABLE ENTERED          6 (UBAR)
MULT. CORR. COEFF.         .83558
STD. ERROR EST.            .12135


                                 ANOVA TABLE

                   DF      SUM OF SQ.          MEAN SQ.           F-RATIO

REGRESSION          1         1.36270           1.36270          92.53208
RESIDUAL           40          .58907            .01473
TOTAL              41         1.95177


                           VARIABLES IN REGRESSION

 VARIABLE     COEFFICIENT          STD.    COMPUTED     BETA     F-OUT   TYP
                                  ERROR    T- VALUE    COEFF.

INTERCEPT      -1.94314
  UBAR   6       .10720          .01114      9.619       .836    9.3E+01(2)


                        VARIABLES NOT IN REGRESSION

 VARIABLE        PART. CORR.   TOLERANCE     F-IN     TYP

TIME/100   1       -.25261         .934      2.7      (2)
APCD OZ.   3        .67087         .569      3.2E+01(2)
SKY CODE   4       -.01817         .874      1.3E-02(2)
 INV/100   5       -.13853         .845      7.6E-01(2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

         APCD
COMPARE [redacted] DATA TO VAN DATA--LOG-LOG TRANS. TEST

PROBLEM NUMBER            1                      F TO ENTER           .01000
STEP NUMBER               2                      F TO REMOVE          .00500
DEPENDENT VARIABLE        2 (FIELD OZ.)          TOLERANCE LEVEL      .00100

VARIABLE ENTERED          3 (APCD OZ.)
MULT. CORR. COEFF.         .91325
STD. ERROR EST.            .09114

                              ANOVA TABLE

                   DF        SUM OF SQ.          MEAN SQ.           F-RATIO

REGRESSION          2          1.62783            .81391           97.98722
RESIDUAL           39           .32395            .00831
TOTAL              41          1.95177


                         VARIABLES IN REGRESSION

VARIABLE     COEFFICIENT           STD.    COMPUTED     BETA     F-OUT   TYP
                                   ERROR   T- VALUE     COEFF.

  INTERCEPT      -.99819
APCD OZ.   3      .50076          .08863     5.650       .489   3.2E+01(2)
     UBAR  6      .06605          .01110     5.952       .515   3.5E+01(2)


                    VARIABLES NOT IN REGRESSION

VARIABLE          PART. CORR.  TOLERANCE     F-IN     TYP

TIME/100   1        -.22019        .917      1.9      (2)
SKY CODE   4         .14280        .845      7.9E-01(2)
 INV/100   5        -.09019        .835      3.1E-01(2)
```

STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE ~~FIELD~~ APCD DATA TO VAN DATA--LOG-LOG TRANS. TEST

```
PROBLEM NUMBER              1                         F TO ENTER            .01000
STEP NUMBER                 3                         F TO REMOVE           .00500
DEPENDENT VARIABLE          2 (FIELD OZ.)             TOLERANCE LEVEL       .00100

VARIABLE ENTERED            1 (TIME/100)
MULT. CORR. COEFF.           .91764
STD. ERROR EST.              .09006


                                  ANOVA TABLE

                      DF        SUM OF SQ.           MEAN SQ.            F-RATIO

REGRESSION             3         1.64353              .54784            67.53836
RESIDUAL              38          .30824              .00811
TOTAL                 41         1.95177


                              VARIABLES IN REGRESSION

VARIABLE          COEFFICIENT            STD.   COMPUTED     BETA      F-OUT    TYP
                                        ERROR   T- VALUE    COEFF.

  INTERCEPT          -.98022
TIME/100   1         -.00501            .00360    -1.391     -.094     1.9      (2)
APCD OZ.   3          .48410            .08840     5.476      .472     3.0E+01(2)
    UBAR   6          .07050            .01142     6.172      .549     3.8E+01(2)


                         VARIABLES NOT IN REGRESSION

VARIABLE           PART. CORR.   TOLERANCE      F-IN     TYP

SKY CODE   4          .14509         .845      8.0E-01(2)
 INV/100   5          .20437         .211      1.6     (2)
```

STEPWISE MULTIPLE LINEAR REGRESSION

COMPARE ~~FIELD~~ APCD DATA TO VAN DATA--LOG-LOG TRANS. TEST

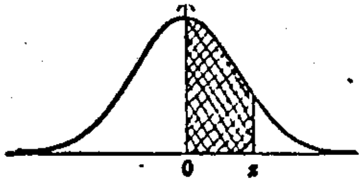
```
PROBLEM NUMBER              1                      F TO ENTER              .01000
STEP NUMBER                 4                      F TO REMOVE             .00500
DEPENDENT VARIABLE          2 (FIELD OZ.)          TOLERANCE LEVEL         .00100

VARIABLE ENTERED            5 (INV/100)
MULT. CORR. COEFF.           .92123
STD. ERROR EST.              .08935

                                    ANOVA TABLE

                       DF        SUM OF SQ.          MEAN SQ.              F-RATIO

REGRESSION              4          1.65641            .41410              51.87380
RESIDUAL               37           .29537            .00798
TOTAL                  41          1.95177


                                VARIABLES IN REGRESSION

     VARIABLE       COEFFICIENT          STD.   COMPUTED     BETA     F-OUT    TYP
                                        ERROR   T- VALUE    COEFF.

   INTERCEPT          -.96411
 TIME/100   1         -.01281           .00711    -1.803     -.240    3.3      (2)
 APCD OZ.   3          .48188           .08772     5.494      .470    3.0E+01(2)
  INV/100   5          .00478           .00377     1.270      .177    1.6      (2)
     UBAR   6          .06654           .01175     5.662      .519    3.2E+01(2)


                          VARIABLES NOT IN REGRESSION

    VARIABLE        PART. CORR.   TOLERANCE     F-IN     TYP

 SKY CODE   4          .06348         .684     1.5E-01(2)
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

         APCD
COMPARE [redacted] DATA TO VAN DATA--LOG-LOG TRANS. TEST


PROBLEM NUMBER            1                    F TO ENTER             .01000
STEP NUMBER               5                    F TO REMOVE            .00500
DEPENDENT VARIABLE        2 (FIELD OZ.)        TOLERANCE LEVEL        .00100

VARIABLE ENTERED          4 (SKY CODE)
MULT. CORR. COEFF.         .92156
STD. ERROR EST.            .09040


                            ANOVA TABLE

                    DF       SUM OF SQ.         MEAN SQ.          F-RATIO

REGRESSION           5        1.65760             .33152          40.56996
RESIDUAL            36         .29418             .00817
TOTAL               41        1.95177


                       VARIABLES IN REGRESSION

VARIABLE     COEFFICIENT            STD.   COMPUTED     BETA     F-OUT   TYP
                                   ERROR   T- VALUE     COEFF.

 INTERCEPT      -.95987
TIME/100   1    -.01165           .00781    -1.493      -.218    2.2      (2)
APCD OZ.   3     .48915           .09077     5.389       .477    2.9E+01(2)
SKY CODE   4     .00514           .01347      .382       .030    1.5E-01(2)
 INV/100   5     .00408           .00424      .963       .151    9.3E-01(2)
    UBAR   6     .06519           .01241     5.253       .508    2.8E+01(2)


F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION.
```

```
STEPWISE MULTIPLE LINEAR REGRESSION

         APCD
COMPARE [redacted]D DATA TO VAN DATA--LOG-LOG TRANS. TEST


                                  SUMMARY TABLE

STEP          VARIABLE          MULT.        STD.          F        NO. IN
NUMBER      NAME   IN  OUT      CORR.        ERROR       RATIO       REG.

  1         UBAR    6          .83558       .12135     92.53208       1
  2      APCD OZ.   3          .91325       .09114     97.98722       2
  3      TIME/100   1          .91764       .09006     67.53836       3
  4       INV/100   5          .92123       .08935     51.87380       4
  5      SKY CODE   4          .92156       .09040     40.56996       5



   RESIDUAL ANALYSIS (YES OR NO)?NO
 ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)?NO
 INPUT NEW DATA FILENAME OR 'STOP'?STOP

   GOOD-BYE.  BE SURE TO DELETE FILE '$REG' BEFORE LEAVING
 THE SYSTEM.
>QUIT
-DEL $REG
-LOGO
 0951  01/09/74
CPU MINS -  0.350
TERMINAL MINS -  28.30
FILE MODULES - 21
----------------
```

STATISTICAL TABLES

TABLE 1. AREAS under the STANDARD NORMAL CURVE from 0 to *z*



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

TABLE 2. Percentiles of the t Distribution*

df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576
df	$-t_{.40}$	$-t_{.30}$	$-t_{.20}$	$-t_{.10}$	$-t_{.05}$	$-t_{.025}$	$-t_{.01}$	$-t_{.005}$

When the table is read from the foot, the tabled values are to be prefixed with a negative sign. Interpolation should be performed using the reciprocals of the degrees of freedom.

*The data of this table extracted from Table III of Fisher and Yates, *Statistical Tables*, with the permission of the authors and publishers, Oliver & Boyd, Ltd., Edinburgh and London.

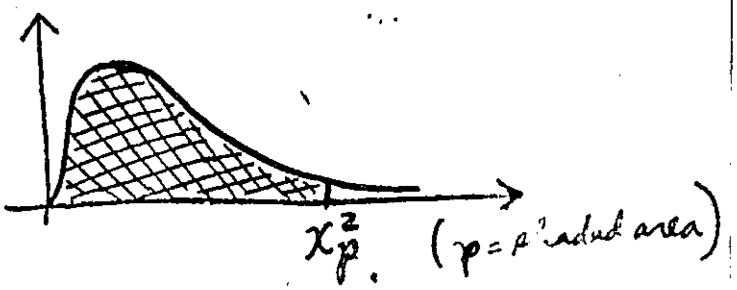
TABLE 3. Percentiles of the χ^2 Distribution

df	Per Cent									
	.5	1	2.5	5	10	90	95	97.5	99	99.5
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

For large values of degrees of freedom the approximate formula

$$\chi_\alpha^2 = n\left(1 - \frac{2}{9n} + z_\alpha\sqrt{\frac{2}{9n}}\right)^3$$

where z_α is the normal deviate and n is the number of degrees of freedom, may be used. For example $\chi_{.99}^2 = 60[1 - .00370 + 2.326(.06086)]^3 = 60(1.1379)^3 = 88.4$ for the 99th percentile for 60 degrees of freedom.



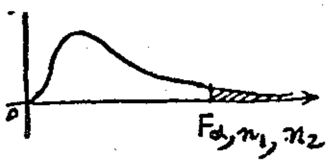


TABLE 4. PERCENTAGE POINTS OF THE F DISTRIBUTION

$\alpha = .05$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	16	20	24	30	50	100	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	252	253	254
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.41	19.43	19.44	19.45	19.46	19.47	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.74	8.69	8.66	8.64	8.62	8.58	8.56	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.84	5.80	5.77	5.74	5.70	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.68	4.60	4.56	4.53	4.50	4.44	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.92	3.87	3.84	3.81	3.75	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.57	3.49	3.44	3.41	3.38	3.32	3.28	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.28	3.20	3.15	3.12	3.08	3.03	2.98	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.07	2.98	2.93	2.90	2.86	2.80	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.91	2.82	2.77	2.74	2.70	2.64	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.79	2.70	2.65	2.61	2.57	2.50	2.45	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.69	2.60	2.54	2.50	2.46	2.40	2.35	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.60	2.51	2.46	2.42	2.38	2.32	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.53	2.44	2.39	2.35	2.31	2.24	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.48	2.39	2.33	2.29	2.25	2.18	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.33	2.28	2.24	2.20	2.13	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.38	2.29	2.23	2.19	2.15	2.08	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.25	2.19	2.15	2.11	2.04	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.31	2.21	2.15	2.11	2.07	2.00	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.28	2.18	2.12	2.08	2.04	1.96	1.90	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.23	2.13	2.07	2.03	1.98	1.91	1.84	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.18	2.09	2.02	1.98	1.94	1.86	1.80	1.73
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.05	1.99	1.95	1.90	1.82	1.76	1.69
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.12	2.02	1.96	1.91	1.87	1.78	1.72	1.65
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.07	1.97	1.91	1.86	1.82	1.74	1.67	1.59
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.03	1.93	1.87	1.82	1.78	1.69	1.62	1.55
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.00	1.90	1.84	1.79	1.74	1.66	1.59	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.81	1.75	1.70	1.65	1.56	1.48	1.39
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.85	1.75	1.68	1.63	1.57	1.48	1.39	1.28
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.80	1.69	1.62	1.57	1.52	1.42	1.32	1.19
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.75	1.64	1.57	1.52	1.46	1.35	1.24	1.00

$\alpha = .01$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	16	20	24	30	50	100	∞
1	4,052	4,999	5,403	5,625	5,764	5,859	5,928	5,981	6,022	6,056	6,106	6,169	6,208	6,234	6,258	6,302	6,334	6,366
2	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.42	99.44	99.45	99.46	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.83	26.69	26.60	26.50	26.35	26.23	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.37	14.15	14.02	13.93	13.83	13.69	13.57	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.89	9.68	9.55	9.47	9.38	9.24	9.13	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.52	7.39	7.31	7.23	7.09	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.47	6.27	6.15	6.07	5.98	5.85	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.67	5.48	5.36	5.28	5.20	5.06	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.11	4.92	4.80	4.73	4.64	4.51	4.41	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.71	4.52	4.41	4.33	4.25	4.12	4.01	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.40	4.21	4.10	4.02	3.94	3.80	3.70	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.16	3.98	3.86	3.78	3.70	3.56	3.46	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.78	3.67	3.59	3.51	3.37	3.27	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.62	3.51	3.43	3.34	3.21	3.11	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.48	3.36	3.29	3.20	3.07	2.97	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.37	3.25	3.18	3.10	2.98	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.45	3.27	3.16	3.08	3.00	2.86	2.76	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.37	3.19	3.07	3.00	2.91	2.78	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.12	3.00	2.92	2.84	2.70	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.23	3.05	2.94	2.86	2.77	2.63	2.53	2.42
22	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.94	2.83	2.75	2.67	2.53	2.42	2.31
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.03	2.85	2.74	2.66	2.58	2.44	2.33	2.21
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	2.96	2.77	2.66	2.58	2.50	2.36	2.25	2.13
28	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.90	2.71	2.60	2.52	2.44	2.30	2.18	2.06
32	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.80	2.62	2.51	2.42	2.34	2.20	2.08	1.96
36	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.72	2.54	2.43	2.35	2.26	2.12	2.00	1.87
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.66	2.49	2.37	2.29	2.20	2.05	1.94	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.32	2.20	2.12	2.03	1.87	1.74	1.60
100	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.36	2.19	2.06	1.98	1.89	1.73	1.59	1.43
200	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.28	2.09	1.97	1.88	1.79	1.62	1.48	1.28
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	1.99	1.87	1.79	1.69	1.52	1.36	1.00

Table 5. Quantiles of the Wilcoxon Signed Ranks Test Statistic

	$w_{.005}$	$w_{.01}$	$w_{.025}$	$w_{.05}$	$w_{.10}$	$w_{.20}$	$w_{.30}$	$w_{.40}$	$w_{.50}$	$\frac{n(n+1)}{2}$
$n = 4$	0	0	0	0	1	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	$\boxed{41}$	48	56	66	73	80	85.5	171
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	82	91	98	105	210

For n larger than 20, the pth quantile w_p of the Wilcoxon signed ranks test statistic may be approximated by $w_p = [n(n+1)/4] + x_p\sqrt{n(n+1)(2n+1)/24}$, where x_p is the pth quantile of a standard normal random variable, obtained from Table 1.

REFERENCE BOOKS ON STATISTICS

I. Parametric Statistics and Regression Analysis

1. Nevill, A. M., Kennedy, J. B., "Basic Statistical Methods for Engineers and Scientists", International Textbook Co., 1964.

2. Kerri, K. D., A. J. Ranzieri, E. Torguson, "Applications of Regression Analysis to Environmental Problems For Highway Projects" Training Manual, Transportation Laboratory, January 1973.

3. Draper, N. R., Smith, "Applied Regression Analysis" John Wiley and Sons, Inc. 1966.

4. Natrella, M. G., "Experimental Statistics" National Bureau of Standards Handbook #91, 1963.

5. Larsen, R. I., "A Mathematical Model For Relating Air Quality To Air Quality Standards", EPA, November 1971.

6. Fong, James, S. L., "Applications of Regression Analysis To Environmental Problems For Highway Projects" January 1973.

II. Nonparametric Statistics

7. Siegel, S., "Nonparametric Statistics for the Behavioral Sciences", McGraw-Hill Book Co., 1956.

8. Conover, W. J., Practical Nonparametric Statistics", John Wiley and Sons, Inc., 1971.

III. Other

9. Beaton, J. L., J. B. Skog, E. C. Shirley and A. J. Ranzieri, "Analysis of Air Quality Data for Highway Projects", Materials and Research Department, Air Quality Manual, CA-HWY-MR657082S-72-13, July 1972.

LISTING OF COMPUTER PROGRAMS

5;LSTAT;CATANOVA

```
1000 DEF DOUBLE FNCHI2(CHIS,V)
1010 DOUBLE CHIS,CHI,C1,C2,PROB,T,QX
1020 INTEGER V,R,D1
1030 IF CHIS<=0 THEN PROB=1 ELSE 1050
1040 GOTO 1160
1050 CHI=SQRT(CHIS), C1=0
1055 IF CHIS>350 THEN ZX=0 ELSE ZX=1/EXP(CHIS/2)
1060 IF V<3 THEN 1110
1070 IF (V MOD 2) THEN C1,C2=CHI, D1=1
      ELSE C1,C2=CHIS/2, D1=2
1080    FOR R=2 TO INT((V-1)/2)
1090    D1=D1+2, C2=C2*CHIS/D1, C1=C1+C2
1100    NEXT R
1110 IF (V MOD 2) THEN 1130 ELSE PROB=ZX*(1+C1)
1120 GOTO 1160
1130 T=1/(1+.2316419*CHI)
1140 QX=T*(.31938153+T*(-.356563782+T*(1.781477937+T*
      (-1.821255978+T*1.330274429))))
1150 PROB=2*ZX*(QX+C1)/SQRT(2*PI)
1160 RETURN PROB
1170 END !

1180 PRINT
1190 PRINT,,"DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION":

1200 INPUT F$
1210 PRINT
1220 IF F$='EXP' THEN 1750
1230 PRINT,,'JOB TITLE':
1240 INPUT HED$
1250 PRINT
1260 PRINT,,'NUMBER OF:'
1270 PRINT,,,'ROWS (CATEGORIES)    ':
1280 INPUT N
1290 PRINT,,,'COLUMNS (GROUPS)     ':
1300 INPUT K
1310 PRINT
1320 PRINT,,'SIGNIFICANCE LEVEL':
1330 INPUT SIG
1340 DIM X(N,K),C(K),CS(K)
1350 OPEN F$,1,INPUT,OLD
1360 MAT INPUT FROM 1:X !

1370 PRINT CHAR(12)
1380 PRINT
1390 PRINT,,'ANOVA FOR CATEGORICAL DATA'
1400 PRINT
1410 PRINT,,HED$
1420 PRINT IN FORM"/6B'DATA:'//10B":
1430    FOR I=1 TO K
1440    PRINT IN FORM'6%':I
1450    NEXT I
1460 F$='6B '+STR(6*K+4)+"('-')//"
1470 PRINT IN FORM F$:
1480    FOR I=1 TO N
1490    PRINT IN FORM"8%'--'":I
1500       FOR J=1 TO K
1510       PRINT IN FORM'6%':X(I,J)
```

```
1520         NEXT J
1530      PRINT
1540      IF NOT (I MOD 10) THEN PRINT
1550      NEXT I
1560      FOR I=1 TO N
1570      NIP=0
1580         FOR J=1 TO K
1590         NIP=NIP+X(I,J), C(J)=C(J)+X(I,J), CS(J)=CS(J)+X(I,J)^2

1600         NEXT J
1610      TSS=TSS+NIP^2, NPP=NPP+NIP
1620      NEXT I !

1630      FOR I=1 TO K
1640      WSS=WSS+CS(I)/C(I)
1650      NEXT I
1660 TSS=NPP/2-TSS/(2*NPP), WSS=(NPP-WSS)/2, BSS=TSS-WSS
167H$CC=R2*NPP-1)*(N-1), DF=(N-1)*(K-1), P=FNCHI2(CC,DF)

1680 PRINT IN FORM"//12B'CATANOVA TABLE'//10B'SOURCE'10B'SS'//
       6B'BETWEEN GROUPS'7%.5% 5B'% EXPLAINED ='7%.5%/":
       BSS,100*R2
1690 PRINT IN FORM"6B'WITHIN  GROUPS'7%.5% 5B'CHI-SQUARE  ='7%.5%/
        38B'DEG.FREEDOM ='7%/15B'TOTAL'7%.5% 5B'PROBABILITY ='
        7%.5%//":WSS,CC,DF,TSS,P
1700 IF P<SIG THEN A$='REJECT' ELSE
        A$='DO NOT REJECT'
1710 PRINT,A$:' THE HYPOTHESIS THAT THE ':K:' GROUP POPULATIONS'
1720 PRINT;;'ARE THE SAME FOR THE ':N:' CATEGORIES.  ':
1730 PRINT IN FORM"'(SIGNIFICANCE LEVEL ='%.3%')'/":SIG
1740 END
1750 PRINT,'***** PROGRAM 5;LSTAT;CATANOVA *****'
1760 PRINT
1770 PRINT;;;'THIS PROGRAM TESTS THE HYPOTHESIS THAT A NUMBER OF'
1780 PRINT;;'SAMPLED GROUP POPULATIONS ARE THE SAME FOR ALL OF A'
179F1810 PRINT;;'R. J. LIGHT AND  B. H. MARGOLIN IN THE JOURNAL OF THE'

1820 PRINT;;'AMERICAN STATISTICAL ASSOCIATION, VOL. 66, NO. 335,'
1830 PRINT;;'SEPTEMBER 1971.'
1840 PRINT
1850 PRINT;;;'BEFORE LINKING THIS PROGRAM, PREPARE A TEXT DATA FILE'

1860 PRINT;;'IN THE FOLLOWING FORMAT:'
1870 PRINT
1880 PRINT,'X(1,1),X(1,2),...,X(1,M)'
1890 PRINT,'X(2,1),X(2,2),...,X(2,M)'
1900 PRINT,'..........................'
1910 PRINT,'X(N,1),X(N,2),...(X(N,M)'
1920 PRINT
1930 PRINT;;;"WHERE X(I,J) = VALUE FOR THE I'TH GROUP OF THE"
1940 PRINT;;;"                          J'TH CATEGORY"
1950 PRINT
1960 PRINT;;;'UPON LINKING THS PROGRAM, THE USER WILL BE'
1970 PRINT;;'REQUESTED TO INPUT THE DATA FILENAME, A JOB TITLE,'
1980 PRINT;;'THE NUMBER OF ROWS (CATEGORIES), THE NUMBER OF COLUMNS'
1990 PRINT;;'(GROUPS) AND THE SIGNIFICANCE LEVEL.'
2000 PRINT
2010 PRINT;;;'OUTPUT CONSISTS OF A LISTING OF THE DATA'
2020 PRINT;;'FOLLOWED BY THE CATANOVA TABLE, THE PERCENTAGE OF'
```

```
2030 PRINT;;'VARIATION EXPLAINED, THE CHI-SQUARE VALUE, DEGREES'
2040 PRINT;;'OF FREEDOM AND ASSOCIATED PROBABILITY.'
2050 PRINT
2060 PRINT;;;'FINALLY, THE CONCLUSION STATEMENT (BASED ON THE'
2070 PRINT;;'SIGNIFICANCE LEVEL) CONCERNING THE NULL HYPOTHESIS WILL'
2080 PRINT;;'BE PRINTED.'
```

5;LSTAT;MPAIR

```
990 DIM X(1000,2),Z(1000),R(1000),S(1000)
1000 PRINT
1010 PRINT,;"DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION":
1020 INPUT F$
1030 IF F$='EXP' THEN 1900
1040 DATA 01 02 04 06 08 11 14 17 21 26 30 36 41 47 54 60 68 75
          83 92 101
1050 DATA -1 01 02 04 06 08 11 14 17 21 25 30 35 40 46 52 59 66
          73 81 90
1060 DATA -1 -1 00 02 03 05 07 10 13 16 20 24 28 33 38 43 49 56
          62 69 77
1070 DATA -1 -1 -1 00 02 03 05 07 10 13 16 19 23 28 32 37 43 49
          55 61 68
1080 DATA .1 .05 .02 .01
1090 MAT READ C(4,5:25),SIG(4)
1095 OPEN F$,1,INPUT,OLD
1100 PRINT,;'JOB TITLE':
1110 INPUT HED$
1120 PRINT
1130 PRINT,;'NUMBER OF:'
1140 PRINT,,;'ROWS (OBSERVATIONS)':
1150 INPUT NM
1160 PRINT,,;'COLUMNS (DISTRIBUTIONS)':
1170 INPUT K
1180 DIM X(NM,K),Z(NM),R(NM),S(NM)
1200 MAT INPUT FROM 1:X !

1210 PRINT CHAR(12)
1220 PRINT
1230 PRINT,;'WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST'
1240 PRINT
1250 PRINT,;HED$
1260 PRINT
1270 PRINT
1280 IF K=2 THEN C1=1, C2=2 ELSE 1300
1290 GOTO 1320
1300 PRINT,;'ENTER THE TWO COLUMN NUMBERS TO BE TESTED':
1310 INPUT C1,C2
1320 N,NP,NN=0
1330    FOR I=1 TO NM
1340    ZZ=X(I,C1)-X(I,C2)
1350    IF ZZ THEN N=N+1, Z(N)=ABS(ZZ), S(N)=SGN(ZZ)
         ELSE 1380
1360    IF S(N)=1 THEN NP=NP+1
1370    IF S(N)=-1 THEN NN=NN+1
1380    NEXT I
1390 PRINT
1400 PRINT,;'OF THE ':NM:' PAIRED OBSERVATIONS, ':NP:' HAD POSITIVE'

1410 PRINT,;'DIFFERENCES AND ':NN:' HAD NEGATIVE DIFFERENCES.'
1420 PRINT
1430 IF N>4 THEN 1450 ELSE PRINT,;
       'UNABLE TO TEST ':N:' PAIRS WITH DIFFERENCES - 5 MINIMUM.'
1440 END
1450 PRINT,;'SIGNIFICANCE LEVEL (2-TAILED)':
1460 IF N<26 THEN PRINT ELSE 1520
1470 PRINT,,;'1 = .10'
1480 IF N>5 THEN PRINT,,;'2 = .05'
```

```
1490 IF N>6 THEN PRINT;;;'3 = ,02'
1500 IF N>7 THEN PRINT;;;'4 = .01'
1510 PRINT;;'INPUT INTEGER SELECTION':
1520 INPUT PROB
1530 PRINT
1540 MAT R=(1)
1550     FOR I=1 TO N-1
1560         FOR J=I+1 TO N
1570         ON SGN(Z(J)-Z(I))+2 GOTO 1600, 1620
158G1600           R(I)=R(I)+
1610         GOTO 1630
1620         R(I)=R(I)+.5, R(J)=R(J)+.5
1630     NEXT J,I
1640 TP,TN=0
1650     FOR I=1 TO N
1660     IF S(I)=1 THEN TP=TP+R(I) ELSE TN=TN+R(I)
1670     NEXT I
1680 T=MIN(TP,TN)
1690 PRINT;;'THE VALUE OF T FOR N = ':N:' IS ':T
1700 PRINT
1710 PRINT
1720 IF N<26 THEN 1810
1730 Q=N*(N+1)/4, QQ=ABS(T-Q)/SQRT(Q*(2*N+1)/6)
1740 T1=1/(1+.2316419*QQ)
1750 QX=T1*(.31938153+T1*(-.356563782+T1*(1.781477937+T1*
       (-1.821255978+T1*1.330274429))))
1760 QX=2*QX/(SQRT(2*PI)*EXP(QQ*QQ/2))
1770 IF QX>PROB THEN A$='DO NOT REJECT' ELSE A$='REJECT'
1780 PRINT IN FORM"6B'ABS(Z) ='2%.5% 3B'PROBABILITY ='%.5%/":QQ,QX
1790 PRINT
1800 GOTO 1850
1810 PRINT;;'CRITICAL VALUE = ':C(PROB,N)
1820 PRINT
1830 IF T>C(PROB,N) THEN A$='DO NOT REJECT' ELSE A$='REJECT'
1840 PROB=SIG(PROB)
1850 PRINT IN FORM"//19%' THE HYPOTHESIS THAT THE TWO DISTRIBUTIONS'
       /6B'ARE FROM THE SAME POPULATION. (SIGNIFICANCE LEVEL ='
       %.3%')'////":A$,PROB
1860 PRINT;;;'ANOTHER ANALYSIS USING THE SAME DATA':
1870 INPUT AN$
1880 IF AN$='YES' THEN 1210
188G1885 IF AN$='YES' THEN 1100
1890 END
1900 PRINT
1910 PRINT,;'***** PROGRAM 5;LSTAT;MPAIR *****'
1920 PRINT
1930 PRINT;;;'THIS PROGRAM TESTS THE HYPOTHESIS THAT TWO SAMPLED'
1940 PRINT;;'DISTRIBUTIONS ARE FROM THE SAME POPULATION BY USE OF'
1950 PRINT;;'THE WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST.  THIS'
1960 PRINT;;"PROCEDURE IS DESCRIBED IN 'NON-PARAMETRIC STATISTICS".

1970 PRINT;;"FOR THE BEHAVIORAL SCIENCES' BY SIDNEY SEIGEL (MCGRAW-"

199F2000 PRINT;;;'BEFORE LINKING THIS PROGRAM, PREPARE A TEXT DATA FILE'

2010 PRINT;;'IN THE FOLLOWING FORMAT:'
2020 PRINT
2030 PRINT;;;'X(1,1),X(1,2),...,X(1,M)'
2040 PRINT;;;'X(2,1),X(2,2),...,X(2,M)'
```

```
2050 PRINT;;;'.........................'
2060 PRINT;;;'X(N,1),X(N,2),...,X(N,M)'
2070 PRINT
2080 PRINT;;"WHERE X(I,J) = THE I'TH OBSERVATION OF THE"
2090 PRINT;;"                         J'TH DISTRIBUTION."
2092 PRINT
2094 PRINT;;'MAX. VALUES - I*J=2000, I=1000'
2100 PRINT
2110 PRINT;;;'UPON LINKING THIS PROGRAM, THE USER WILL BE'
2120 PRINT;;'REQUESTED TO INPUT THE DATA FILENAME, A JOB TITLE,'
2130 PRINT;;'THE NUMBER OF ROWS (OBSERVATIONS), AND THE NUMBER'
2140 PRINT;;'OF COLUMNS (DISTRIBUTIONS) IN THE DATA FILE.'
2150 PRINT
2160 PRINT;;;'*****FOR EACH PROBLEM *****'
2170 PRINT
2180 PRINT;;;'IF THE DATA FILE CONTAINS MORE THAN TWO DISTRI-'
2190 PRINT;;'BUTIONS, THE USER WILL BE REQUESTED TO INPUT THE TWO'
2200 PRINT;;'COLUMN NUMBERS HE WISHES TO TEST.'
2210 PRINT
2220 PRINT;;;'THE NUMBER OF POSITIVE AND NEGATIVE DIFFERENCES WILL'
2230 PRINT;;'THEN BE OUTPUT AND THE USER WILL THEN BE REQUESTED TO'
2240 PRINT;;'INPUT (BY INTEGER SELECTION IF N<26 OR OTHERWISE ACTUAL'
2250 PRINT;;'VALUE) THE TWO-TAILED SIGNIFICANCE LEVEL.  (IF A ONE-'
2260 PRINT;;'TAILED TEST, ENTER A VALUE EQUAL TO TWICE THE SIGNIFI'
2270 PRINT;;'CANCE LEVEL DESIRED)'
2280 PRINT
2290 PRINT;;;'THE OUTPUT THEN CONSISTS OF THE VALUE OF T, EITHER'
2300 PRINT;;'THE CRITICAL VALUE IF N<26 OR THE ABSOLUTE VALUE OF Z'
2310 PRINT;;'AND THE CALCULATED PROBABILITY IF N>25, AND THE STATE'
2320 PRINT;;'MENT OF WHETHER TO ACCEPT OR REJECT THE HYPOTHESIS THAT'
2330 PRINT;;'THE TWO DISTRIBUTIONS ARE FROM THE SAME POPULATION.'
2340 PRINT
2350 PRINT;;;'THE USER IS THEN REQUESTED TO INPUT A "YES" OR "NO"'
2360 PRINT;;'TO WHETHER HE WISHES ANOTHER ANALYSIS USING THE SAME'
2370 PRINT;;'DATA.  IF "YES" THE PROGRAM WILL START WITH A NEW'
2380 PRINT;;'ANALYSIS OF THE SAME DATA.'
2390 PRINT
2400 PRINT;;;'IF "NO" IS INPUT, THE USER WILL BE ASKED IF THERE IS'
2410 PRINT;;'ANOTHER PROBLEM IN THE DATA FILE.  IF "YES" IS ENTERED,'
2420 PRINT;;'THE PROGRAM WILL START BY REQUESTING A NEW JOB TITLE.'
```

5;LSTAT;FRIED

```
1000 DEF DOUBLE FNCHI2(CHIS,V)
1010 DOUBLE CHIS,CHI,C1,C2,PROB,T,QX
1020 INTEGER V,R1,D1
1030 IF CHIS<=0 THEN PROB=1 ELSE 1050
1040 GOTO 1170
1050 CHI=SQRT(CHIS), C1=0
1060 IF CHIS>350 THEN ZX=0 ELSE ZX=1/EXP(CHIS/2)
1070 IF V<3 THEN 1120
1080 IF (V MOD 2) THEN C1,C2=CHI, D1=1      ELSE C1,C2=CHIS/2, D1=2
1090 FOR R1=2 TO INT((V-1)/2)
1100 D1=D1+2, C2=C2*CHIS/D1, C1=C1+C2
1110 NEXT R1
1120 IF (V MOD 2) THEN 1140 ELSE PROB=ZX*(1+C1)
1130 GOTO 1170
1140 T=1/(1+.2316419*CHI)
1150 QX=T*(.31938153+T*(-.356563782+T*(1.781477937+T*      (-1.821255978+T
    *1.330274429))))
1160 PROB=2*ZX*(QX+C1)/SQRT(2*PI)
1170 RETURN PROB
1180 END !
1190 PRINT ;;'MODIFIED  FEB. 74'
1195 PRINT
1200 PRINT;;"DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION":
1210 INPUT F$
1220 PRINT
1230 IF F$='EXP' THEN LINK '5;LSTAT;FRIEDEXP' !
1240 INTEGER ROWS,COLS,REPS,NC
1250 OPEN F$,1,INPUT,OLD
1260 PRINT;;'JOB TITLE':
1270 INPUT HED$
1280 PRINT
1290 PRINT;;'NUMBER OF:'
1300 PRINT;;;'OBSERVATIONS / REPLICATION':
1310 INPUT ROWS
1320 PRINT;;;'COLUMNS (TREATMENTS)':
1330 INPUT COLS
1340 PRINT;;;'REPLICATIONS        ':
1350 INPUT REPS
1360 IF ROWS>0 AND COLS>0 AND  REPS>0 THEN 1390 ELSE PRINT
1370 PRINT;;'ROWS,COLUMNS AND REPLICATIONS MUST BE 1 OR GREATER.'
1380 GOTO 1280 !
1390 N=COLS*REPS
1400 INTEGER COL(COLS)
1410 DIM X(REPS,ROWS,COLS),Y(REPS,ROWS,COLS),A(N),R(N),SUM(N)
1420 MAT INPUT FROM 1:X !
1421 CLOSE 1
1422 PRINT "DO YOU WANT TO SKIP INTERMEDIATE PRINT AND STORE DATA ON A FIL
    E";
1423 INPUT SKP$
1424 IF LEFT (SKP$,1)#'Y' THEN 1435
1425 PRINT "ENTER DATA FILE NAME , OLD OR NEW";
1426 INPUT FIL$,AG$
1427 IF AG$='OLD' THEN 1430
1428 OPEN FIL$,1,SYMBOLIC,OUTPUT,SEQUENTIAL,NEW
1429 GOTO 1431
```

```
1430 OPEN FIL$,1,SYMBOLIC,SEQUENTIAL,OUTPUT,OLD
1431 X$=STR(COLS)+'(4%.% B)/'
1432 MAT PRINT ON 1 IN FORM X$:X
1433 GOTO 1670
1435 PRINT CHAR(12)
1440 PRINT
1450 PRINT;;'FRIEDMAN 2-WAY ANOVA BY RANKS'
1460 PRINT
1470 PRINT;;HED$
1480 PRINT IN FORM"/6B'DATA:'//10B":
1490 FOR I=1 TO COLS
1500 PRINT IN FORM'6%':I
1510 NEXT I
1520 F$='6B'+STR(6*COLS+4)+"('-')//"
1530 PRINT IN FORM F$:
1540 FOR I=1 TO REPS
1550 PRINT,;' REPLICATION ':I
1560 FOR J=1 TO ROWS
1570 PRINT IN FORM"8%'--'":J
1580 FOR K=1 TO COLS
1590 PRINT IN FORM'4%.%':X(I,J,K)
1600 NEXT K
1610 PRINT
1620 NEXT J
1630 PRINT
1640 NEXT I !
1650 DO 1435: 1470
1660 PRINT
1670 PRINT;;'INPUT NUMBER OF COLUMNS IN ANALYSIS':
1680 INPUT NC
1690 IF NC>2 THEN 1720 !
1700 IF NC=2 THEN PRINT;;;        'UNABLE TO COMPARE ':NC:' OBSERVATIONS.' E
   LSE PRINT;;;        'LINK 5;LSTAT;MPAIR FOR TEST OF TWO DISTRIBUTIONS.'
1710 GOTO 1660 !
1720 IF NC>COLS THEN 2560
1730 INTEGER COL(NC)
1740 IF NC<COLS THEN 1790 !
1750 FOR I=1 TO NC
1760 COL(I)=I
1770 NEXT I
1780 GOTO 1920
1790 PRINT;;'INPUT ':NC:' COLUMN NUMBERS IN ASCENDING ORDER'
1800 PRINT;;
1810 MAT INPUT COL
1820 IF COL(1)>0 THEN 1850 ELSE PRINT
1830 PRINT;;'INPUT COLUMN NUMBERS FROM 1 TO ':COLS:' ONLY'
1840 GOTO 1670 !
1850 FOR I=2 TO NC
1860 IF COL(I)>COL(I-1) THEN 1890 ELSE PRINT
1870 PRINT;;'COLUMN NUMBERS ARE NOT IN ASCENDING ORDER'
1880 GOTO 1670
1890 NEXT I !
1900 IF COL(NC)>COLS THEN PRINT ELSE 1920
1910 GOTO 1830
1920 N=NC*REPS, F$='6B'+STR(6*NC+4)+"('-')//"
1921 IF LEFT(SKP$,1)='Y' THEN 1980
1930 PRINT IN FORM"/6B'RANKS'//10B":
1940 FOR I=1 TO NC
1950 PRINT IN FORM'6%':COL(I)
1960 NEXT I
```

```
1970 PRINT IN FORM F$:
1980 DIM A(N),R(N),Y(REPS,ROWS,NC),SUM(NC)
1990 MAT SUM=(0)
2000 FOR I=1 TO ROWS
2010 L=0
2020 FOR J=1 TO NC
2030 FOR K=1 TO REPS
2040 L=L+1, A(L)=X(K,I,COL(J))
2050 NEXT K,J
2060 MAT R=(0)
2070 FOR J=1 TO N
2080 IF R(J) THEN 2190 ELSE SM,EQ=0, Q=A(J)
2090 FOR K=1 TO N
2100 IF A(K)>Q THEN 2120
2110 IF A(K)<Q THEN SM=SM+1 ELSE EQ=EQ+1, R(K)=-1
2120 NEXT K
2130 IF EQ>1 THEN 2150 ELSE R(J)=SM+1
2140 GOTO 2190
2150 Q=SM+.5*(EQ+1)
2160 FOR K=1 TO N
2170 IF R(K)=-1 THEN R(K)=Q
2180 NEXT K
2190 NEXT J
2200 L=0
2210 FOR J=1 TO NC
2220 FOR K=1 TO REPS
2230 L=L+1, Y(K,I,J)=R(L), SUM(J)=SUM(J)+R(L)
2240 NEXT K,J,I !
2242 IF LEFT(SKP$,1)#'Y' THEN 2250
2243 Y$=STR(NC)+'(4%.% B)/'
2244 C$="/"+Y$
2245 MAT PRINT ON 1 IN FORM C$:COL
2246 MAT PRINT ON 1 IN FORM Y$:Y
2247 GOTO 2360
2250 FOR I=1 TO REPS
2260 PRINT,;' REPLICATION ':I
2270 FOR J=1 TO ROWS
2280 PRINT IN FORM"8%'--'":J
2290 FOR K=1 TO NC
2300 PRINT IN FORM'4%.%':Y(I,J,K)
2310 NEXT K
2320 PRINT
2330 NEXT J
2340 PRINT
2350 NEXT I !
2360 DO 1435: 1470
2370 PRINT
2380 PRINT,;'COLUMN RANK TOTALS:'
2390 PRINT
2400 S=0, K=.5*(ROWS*REPS*(REPS*NC+1))
2410 FOR I=1 TO NC
2420 S=S+(SUM(I)-K)^2
2430 PRINT IN FORM"9B'COLUMN'3%' ='5%.%/":COL(I),SUM(I)
2440 NEXT I
2445 IF LEFT(SKP$,1)#'Y' THEN 2450
2446 MAT PRINT ON 1 IN FORM C$:SUM
2450 CHI=6*S/(NC*REPS*K), P=FNCHI2(CHI,NC-1)
2460 PRINT
2470 PRINT,;'SIGNIFICANCE LEVEL':
2480 INPUT PROB
```

```
2490 IF (NC=3 AND N<10)        OR (NC=4 AND N<05)        THEN PRINT ELSE 2510
2500 PRINT;;'INEXACT PROBABILITY - REFER TO PROGRAM INFORMATION.'
2510 IF P>PROB THEN 2525
2520 PRINT IN FORM"//6B'CHI-SQUARE ='5%.3% 5B'DF ='4% 5B'PROBABILTTY ='
       %.5%//'REJECT HYPOTHESIS THAT THE'3%' DISTRIBUTIONS ARE'/        6B
        ROM THE SAME POPULATION.'///":CHI,NC-1,P,A$,NC
2522 GOTO 2530
2525 PRINT IN FORM"//6B'CHI-SQUARE ='5%.3% 5B'DF ='4% 5B'PROBABILITY ='
       %.5%//'DO NOT REJECT HYPOTHESIS THAT THE'3%' DISTRIBUTTONS ARE'/
            6B'FROM THE SAME POPULATION.'///":CHI,NC-1,P,A$,NC
2530 PRINT;;'ANOTHER ANALYSIS USING THE SAME DATA';
2540 INPUT AN$
2550 IF AN$='YES' THEN 1650
2560 IF LEFT(SKP$,1)#'Y' THEN 2570
2565 CLOSE 1
2570 END
```

5;LSTAT;CONTIN

```
100 DIM A(30,30),TR(30),TC(30)
110 PRINT"DATA FILENAME OR 'EXP' FOR INPUT FORMAT":
120 INPUT F$
130 IF F$#'EXP' THEN 1000
135 PRINT
140 PRINT
141 PRINT'                    ***** PROGRAM 5;LSTAT;CONTIN *****'
142 PRINT
150 PRINT;;'THIS PROGRAM COMPUTES CHI-SQUARE FOR CONTINGENCY TABLES'
160 PRINT;'OF UP TO 30 ROWS AND 30 COLUMNS.'
170 PRINT
180 PRINT;'DATA IS ENTERED INTO A SEQUENTIAL TEXT FILE AS FOLLOWS:'
190 PRINT
200 PRINT;;'TIT$,NR,NC'
290 PRINT;;'X(1,1),X(1,2),....,X(1,NC)'
300 PRINT;;'X(2,1),X(2,2),....,X(2,NC)'
310 PRINT;;'..........................'
320 PRINT;;'X(NR,1),X(NR,2),....,X(NR,NC)'
330 PRINT
340 PRINT;'WHERE:'
350 PRINT;;'TIT$=JOB TITLE'
360 PRINT;;'NR =NUMBER OF ROWS 1<NR<31'
370 PRINT;;'NC =NUMBER OF COLUMNS 1<NC<31'
380 PRINT;;'X(I,J)= DATA'
390 PRINT
400 PRINT;;'OUTPUT CONSISTS OF ACTUAL AND EXPECTED FREQUENCIES FOR EAC
H'
410 PRINT;'CELL AND THE CHI-SQUARE (WITH ASSOCIATED DEGREES OF FREE-'
420 PRINT;'DOM).'
430 PRINT
440 PRINT;;'UPON COMPLETION, THE PROGRAM WILL CHECK TO SEE IF ANOTHER'
450 PRINT;'SET OF DATA IS IN THE FILE FOR ANALYSIS.  IF ANOTHER SET'
460 PRINT;'IN THE SAME FORMAT IS AVAILABLE, AN ANALYSIS ON THE NEW'
470 PRINT;'DATA WILL BE PERFORMED.'
480 PRINT
490 PRINT;;'THE PROGRAM WILL CONTINUE UNTIL AN END OF FILE CONDITION'
500 PRINT;'IS ENCOUNTERED.'
510 PRINT
520 PRINT;;'LIST 5;LSTAT;CONFIL FOR A TYPICAL DATA FILE.'
530 END
1000 OPEN F$,1,INPUT,OLD
1010 ON ENDFILE (1) GOTO 1380
1020 PRINT CHAR(12)
1030 SLEEP 2
1040 PRINT
1050 PRINT
1060 INPUT FROM 1:TIT$,N,M
1070 PRINT;'CHI-SQUARE CONTINGENCY TABLE...':TIT$
1080 PRINT
1090 DIM A(N,M)
1100 MAT INPUT FROM 1:A
1110 NDF=(N-1)*(M-1)                        ! NDF=DEGREES OF FREEDOM
1120 IF NDF THEN 1150
1130 PRINT;'DEGREES OF FREEDOM=0.  PROBLEM NOT EXECUTABLE.'
1140 GOTO 1020
1150 CS,GT=0
1160 MAT TR=(0)
1170 MAT TC=(0)
1180 FOR I=1 TO N
```

```
1190 FOR J=1 TO M
1200 TR(I)=TR(I)+A(I,J)                          ! ROW TOTALS
1210 TC(J)=TC(J)+A(I,J)                          ! COLUMN TOTALS
1220 GT=GT+A(I,J)                                ! GRAND TOTAL
1230 NEXT J,I
1240 PRINT,'CELL','ACTUAL','EXPECTED'
1250 PRINT
1260 FOR I=1 TO N
1270 FOR J=1 TO M
1280 E=TR(I)*TC(J)/GT
1290 PRINT,I:',':J,A(I,J),E:
1300 IF E<1 THEN PRINT'*' ELSE PRINT
1310 KFAC=A(I,J)-E
1320 IF NDF=1 THEN KFAC=ABS(KFAC)-.5
1330 CS=CS+KFAC*KFAC/E
1340 NEXT J,I
1350 PRINT
1360 PRINT,'CHI-SQUARE (':NDF:' D.F.) = ':CS
1370 GOTO 1030
1380 END
```

5;LSTAT;LINREG

```
100 DIM X(1000),Y(1000),T(10)
110 PRINT
120 PRINT,'***** MODIFIED 12/6/72 *****'
130 PRINT
140 PRINT;'INPUT DATA FILENAME OR "EXP" FOR PROGRAM EXPLANATION':

150 INPUT F$
160 IF F$#'EXP' THEN 660
170 PRINT
180 PRINT
190 PRINT,'***** PROGRAM 5;LSTAT;LINREG *****'
200 PRINT
210 PRINT;;'THIS PROGRAM PERFORMS A LEAST SQUARES LINEAR REGRESSION'

220 PRINT;'ANALYSIS ON UP TO 1000 PAIRS OF BIVARIATE DATA.  EITHER'

230 PRINT;'THE INDEPENDENT(X) VALUES OR THE DEPENDENT(Y) VALUES OR'

240 PRINT;'BOTH MAY BE TRANSFORMED TO LOG10 OR INVERSE VALUES.'
250 PRINT;'(TYPE OF TRANSFORMATION DESIRED WILL BE REQUESTED BY INPUT)
'
260 PRINT
270 PRINT;;'DATA IS READ INTO THE PROGRAM FROM A SEQUENTIAL TEXT'

280 PRINT;'FILE IN EITHER OF THE FOLLOWING FORMATS (FORMAT TYPE'

290 PRINT;'ALSO REQUESTED BY INPUT.)'
300 PRINT
310 PRINT;;'EITHER'
320 PRINT;;;'TIT$,N'
330 PRINT;;;'X(1),X(2),....,X(N)'
340 PRINT;;;'Y(1),Y(2),....,Y(N)'
350 PRINT
360 PRINT;;'OR:'
370 PRINT;;;'TIT$,N'
380 PRINT;;;'X(1),Y(1),X(2),Y(2),....,X(N),Y(N)'
390 PRINT
400 PRINT;'WHERE:'
410 PRINT;;'TIT$ = JOB TITLE'
420 PRINT;;'   N = NUMBER OF DATA PAIRS'
430 PRINT;;'X(I) = INDEPENDENT VARIABLE DATA'
440 PRINT;;'Y(I) = CORRESPONDING DEPENDENT VARIABLE DATA'

450 PRINT
460 PRINT;;'OUTPUT CONSISTS OF:'
470 PRINT;;;'1. THE REGRESSION EQUATION.'
480 PRINT;;;'2. THE STANDARD ERROR OF REGRESSION.'
490 PRINT;;;'3. THE INDEX OF DETERMINATION.'
500 PRINT;;;'4. THE COEFFICIENT OF CORRELATION.'
510 PRINT;;;'5. AN ANALYSIS OF VARIANCE TABLE.'
520 PRINT;;;'6. THE MEAN, VARIANCE AND STANDARD DEVIATION FOR'
530 PRINT;;;;'   BOTH (TRANSFORMED) VARIABLES.'
540 PRINT;;;'7. THE STANDARD ERROR AND 95% CONFIDENCE LIMITS FOR'
550 PRINT;;;;'      BOTH REGRESSION COEFFICIENTS.'
560 PRINT
570 PRINT;;'YOU WILL THEN BE ASKED IF YOU WISH A LISTING OF VALUES.'

580 PRINT;'IN ANSWER, YOU WILL BE REQUESTED TO INPUT:'
```

```
590 DO 870: 900
600 PRINT
610 PRINT;;'ADDITIONAL ANALYSES WILL BE PERFORMED IF SUBSEQUENT DATA'
620 PRINT;'IN THE FILE IS IN THE PRESCRIBED FORMAT.'
630 PRINT
640 PRINT;;'LIST 5;LSTAT;LINFIL FOR A TYPICAL DATA FILE.'
650 END
660 PRINT
670 PRINT;'DATA TRANSFORMATION CODES:'
680 PRINT;;'0 - NO TRANSFORMATION OF DATA'
690 PRINT;;'1 - V = LOG10(V)'
700 PRINT;;'2 - V = 1/V'
710 PRINT;'ENTER TRANSFORMATION CODES FOR X,Y':
720 INPUT TRX,TRY
730 PRINT
740 IF TRX>=0 AND TRY>=0 AND TRX<3 AND TRY<3 THEN 770
750 PRINT;'INPUT TWO CODES BETWEEN 0 AND 2':
760 GOTO 720
770 PRINT;'FILE FORMAT:'
780 PRINT;;'1 = ALL X VALUES, THEN ALL Y VALUES'
790 PRINT;;'2 = XY PAIRS'
800 PRINT;;'WHICH':
810 INPUT TYP
820 PRINT
830 IF TYP=1 OR TYP=2 THEN 860
840 PRINT;'ENTER EITHER 1 OR 2':
850 GOTO 810
860 PRINT;'WHEN ASKED "WHAT NEXT?" ENTER:'
870 PRINT;;'0  FOR NO LISTING OF X, Y-ACTUAL AND Y-CALCULATED'
880 PRINT;;'1  FOR LISTING WITH DIFFERENCES AND % DIFFERENCES'
890 PRINT;;'2  FOR LISTING WITH 95% CONFIDENCE LIMITS OF YBAR'
900 PRINT;;'3  FOR LISTING WITH 95% PREDICTION LIMITS FOR Y'
910 DATA 1.95996,2.37226,2.82248,2.55582,4.0625
920 DATA 12.706,4.303,3.182,2.776,2.571,2.447,2.365,2.306,
       2.262,2.228
930 READ A0,A1,A2,A3,A4
940 MAT READ T
950 DEF FNT(X)
960 REAL X
970 X=1/X
980 RETURN A0+X*(A1+X*(A2+X*(A3+X*A4)))
990 END
1000 OPEN F$,1,INPUT,OLD
1010 ON ENDFILE(1) GOTO 2160
1020 PRINT CHAR(12)
1030 PRINT
1040 INPUT FROM 1:TIT$,N
1050 IF N>1000 THEN PRINT'TOO MUCH DATA - 1000 PAIRS MAX.' ELSE 1070
1060 END
1070 DIM X(N),Y(N)
1080 PRINT;'LEAST SQUARES LINEAR REGRESSION ANALYSIS'
1090 PRINT
```

```
1100 PRINT;TIT$
1110 PRINT
1120 PRINT
1130 IF TYP=2 THEN 1150
1140 MAT INPUT FROM 1:X,Y
1150 MX,MY,SX,SY,SC=0
1160     FOR I=1 TO N
1170     K=(I-1)/I
1180     IF TYP=2 THEN INPUT FROM 1:X(I),Y(I)
1190     IF TRX=1 THEN X(I)=LOG10(X(I))
1200     IF TRX=2 THEN X(I)=1/X(I)
1210     IF TRY=1 THEN Y(I)=LOG10(Y(I))
1220     IF TRY=2 THEN Y(I)=1/Y(I)
1230     KX=X(I)-MX
1240     KY=Y(I)-MY
1250     MX=MX+KX/I
1260     MY=MY+KY/I
1270     SX=SX+K*KX*KX
1280     SY=SY+K*KY*KY
1290     SC=SC+K*KX*KY
1300     NEXT I
1310 SLOP=SC/SX
1320 YINT=MY-SLOP*MX
1330 RSQ=SLOP*SC/SY
1340 SSE=SY-SLOP*SC
1350 SSR=SY-SSE
1360 MSE=SSE/(N-2)
1370 SA=SQRT(MSE*(1/N+MX*MX/SX))
1380 SB=SQRT(MSE/SX)
1390 IF TRX=0 THEN X2$='X', X1$='*X'
1400 IF TRX=1 THEN X2$='LOG10(X)', X1$='*'+X2$
1410 IF TRX=2 THEN X2$='1/X', X1$='/X'
1420 IF TRY=0 THEN Y$='Y'
1430 IF TRY=1 THEN Y$='LOG10(Y)'
1440 IF TRY=2 THEN Y$='1/Y'
1450 PRINT;'REGRESSION EQUATION: ':Y$:'=':YINT:
1460 IF SLOP>=0 THEN PRINT'+':
1470 PRINT SLOP:X1$
1480 PRINT
1490 PRINT;'NUMBER OF OBSERVATIONS':TAB(30):'= ':N
1500 PRINT;'STD. ERROR OF ':Y$:' ON X':TAB(30):'= ':SQRT(MSE)
1510 PRINT;'INDEX OF DETERMINATION':TAB(30):'= ':RSQ
1520 PRINT;'COEFF. OF CORRELATION':TAB(30):'= ':SGN(SLOP)*SQRT(RSQ)

1530 PRINT
1540 PRINT
1550 PRINT;'SOURCE',,'  SS','DF','  MS','  F'
1560 PRINT
1570 PRINT;'REGRESSION',SSR,1,SSR,SSR/MSE
1580 PRINT;'RESIDUAL',,SSE,N-2,MSE
1590 PRINT
1600 PRINT;'TOTAL',,SY,N-1
1610 PRINT
1620 PRINT
1630 PRINT;'VAR.','MEAN','VARIANCE','STD. DEV.'
1640 PRINT
1650 PRINT IN FORM'8% 4B':X2$
1660 PRINT MX,SX/(N-1),SQRT(SX/(N-1))
1670 PRINT IN FORM'8% 4B':Y$
```

```
1680 PRINT MY,SY/(N-1),SQRT(SY/(N-1))
1690 PRINT
1700 PRINT
1710 PRINT,'STD. ERROR','95% CONFIDENCE LIMITS'
1720 IF N>12 THEN CL=FNT(N-2) ELSE CL=T(N-2)
1730 PRINT
1740 S1=SA, S2=YINT-CL*SA, S3=YINT+CL*SA
1750 IF TRY=1 THEN S2=10^S2, S3=10^S3
1760 IF TRY=2 THEN S2=1/S2, S3=1/S3 ELSE 1768
1762 IF SGN(YINT)=SGN(S2) THEN S3$=STR(S2) ELSE S3$='?'
1764 IF SGN(YINT)=SGN(S3) THEN S2$=STR(S3) ELSE S2$='?'
1766 GOTO 1770
1768 S2$=STR(S2), S3$=STR(S3)
1770 PRINT;'YINTCPT',S1,S2$,S3$
1780 PRINT;'SLOPE',SB,SLOP-CL*SB,SLOP+CL*SB
1790 PRINT
1800 PRINT
1810 PRINT;'WHAT NEXT':
1820 INPUT AN
1830 IF AN=0 THEN 1020
1840 DO 1020,1030,1080:1120,1450:1480
1850 PRINT,'X-ACTUAL','Y-ACTUAL','Y-CALC',
1860 ON AN GOTO 1910, 1890
1870 PRINT'95% PREDICTION LIMITS'
1880 GOTO 1920
1890 PRINT'95% CONFIDENCE LIMITS'
1900 GOTO 1920
1910 PRINT'DIFFERENCE','PERCENT'
1920 PRINT
1930    FOR I=1 TO N
1940    YC=YINT+SLOP*X(I)
1950    X1=X(I)
1960    IF TRX=1 THEN X1=10^X1
1970    IF TRX=2 THEN X1=1/X1
1980    Y1=Y(I), Y2=YC
1990    IF TRY=1 THEN Y1=10^Y1, Y2=10^Y2
2000    IF TRY=2 THEN Y1=1/Y1, Y2=1/Y2
2010    PRINT,X1,Y1,Y2,
2020    ON AN GOSUB 2050, 2080, 2100
2030    NEXT I
2040 GOTO 1790
2050 DIF=Y1-Y2
2060 PRINT DIF,100*DIF/Y2
2070 RETURN
2080 L=SQRT(MSE*(1/N+(X(I)-MX)^2/SX))*CL
2090 GOTO 2110
2100 L=SQRT(MSE*(1+1/N+(X(I)-MX)^2/SX))*CL
2110 Y1=YC-L, Y2=YC+L
2120 IF TRY=1 THEN Y1=10^Y1, Y2=10^Y2
2130 IF TRY=2 THEN Y1=1/Y1, Y2=1/Y2 ELSE 2138
2132 IF SGN(YC)=SGN(Y1) THEN Y2$=STR(Y1) ELSE Y2$='?'
2134 IF SGN(YC)=SGN(Y2) THEN Y1$=STR(Y2) ELSE Y1$='?'
2136 GOTO 2140
2138 Y1$=STR(Y1), Y2$=STR(Y2)
2140 PRINT Y1$,Y2$
2150 RETURN
2160 END
```

PROGRAM LISTING FOR

5;LSTAT;STPREG

```
COPY %STPREG TO TEL TEXT


800 PRINT
900 PRINT,'***** MODIFIED 2/27/73 *****'
950 PRINT
1000 DOUBLE Q
1010 PRINT;"DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION":
1020 INPUT F$
1030 PRINT
1040 IF F$='EXP' THEN 1140
1050 PRINT;;"A WORK FILE '$REG' HAS BEEN CREATED FOR USE BY THIS"
1060 PRINT;'PROGRAM.  DELETE THIS FILE WHEN COMPLETED WITH THIS RUN.'

1065 PRINT
1070 PRINT;'JOB TITLE':
1080 INPUT HED$
1090 HED$="'"+CHAR(12)+"'//6B'STEPWISE MULTIPLE LINEAR REGRESSION'"
      +"///6B'"+HED$+"'///"
1100 Q=2
1110 OPEN '$REG',1,RANDOM,BINARY
1120 PRINT ON 1 AT 1:HED$,F$,Q
1130 LINK '5;LSTAT;CORRE'
1140 PRINT CHAR(12)
1150 PRINT
1160 PRINT,'***** PROGRAM 5;LSTAT;STPREG *****'
1170 PRINT
1180 PRINT;;;'THIS STEPWISE MULTIPLE LINEAR REGRESSION ANALYSIS IS A'
1190 PRINT;;'SYSTEM OF SIX BASIC BINARY PROGRAMS WHICH ARE LINKED IN A'
1200 PRINT;;'MANNER WHICH ALLOWS THE USER SOME CONTROL OF THE DEPTH OF'
1210 PRINT;;'ANALYSIS.  THE OUTPUT IS SIMILAR TO THE UNIV. OF CALIF.'
1220 PRINT;;'BMD02R OR CSD STS010 COMPUTER PROGRAMS.'
1230 PRINT
1240 PRINT;;;'PRIOR TO LINKING THIS PROGRAM, PREPARE A DATA FILE IN'
1250 PRINT;;'THE FOLLOWING FORMAT:'
1260 PRINT
1270 PRINT,'X(1,1),X(1,2),...,X(1,M)'
1280 PRINT,'X(2,1),X(2,2),...,X(2,M)'
1290 PRINT,'.......................'
1300 PRINT,'X(N,1),X(N,2),...,X(N,M)'
1310 PRINT
1320 PRINT;;"WHERE X(I,J) = DATA FOR THE I'TH OBSERVATION AND THE"
1330 PRINT;;"                   J'TH VARIABLE."
1340 PRINT
1350 PRINT;;;'UPON LINKING 5;LSTAT;STPREG, YOU WILL BE REQUESTED TO'
1360 PRINT;;'INPUT THE DATA FILE NAME.  AT THIS TIME, A BINARY WORK'
1370 PRINT;;"FILE, '$REG', WILL BE CREATED WHICH MUST BE DELETED (IN"
1380 PRINT;;'THE EXECUTIVE) BEFORE LINKING THIS PROGRAM (OR ANY REG-'
1390 PRINT;;'RESSION PROGRAM IN 5;LSTAT WHEN CONVERSION IS COMPLETE)'
1400 PRINT;;'AGAIN.  AN EXCEPTION TO THIS WILL BE EXPLAINED LATER.'
1410 PRINT
1420 PRINT;;;'ALTHOUGH LINKING TO THE VARIOUS PROGRAMS IS AUTOMATIC,'
1430 PRINT;;'THE INPUT REQUIRED AND THE OUTPUT FROM EACH PROGRAM WILL'
1440 PRINT;;'BE EXPLAINED.  THIS WILL ALLOW THE USER TO ENTER AT THE'
1450 PRINT;;'PROPER POINT IF AN ANALYSIS IS PREMATURELY ABORTED.'
1460 PRINT
1470 PRINT;;;'FOUR OF THE SIX PROGRAMS ARE (OR WILL BE) COMMON TO ALL'
1480 PRINT;;'5;LSTAT REGRESSION SYSTEMS.  EXPLANATIONS OF THESE FOUR'
1490 PRINT;;'PROGRAMS MAY BE OBTAINED BY THE EXECUTIVE COMMAND:'
1500 PRINT
```

```
1510 PRINT,"'COPY 5;LSTAT;REGEXP TO TEL TEXT'"
1515 PRINT
1520 PRINT
1530 PRINT;;'THE SEQUENCE OF PROGRAMS ARE EXPLAINED BELOW:'
1540 PRINT
1550 PRINT;;'5;LSTAT;STPREG'
1560 PRINT
1570 PRINT;;'THIS PROGRAM REQUESTS THE DATA FILENAME AND JOB TITLE.'
1580 PRINT;;"IT OPENS '$REG' AND ENTERS THE FILE NAME, HEADING AND"
1590 PRINT;;'AN IDENTIFICATION THAT THIS IS STPREG INTO THE WORK FILE.'
1600 PRINT;;'IT THEN LINKS TO 5;LSTAT;CORRE.'
1610 PRINT
1620 PRINT;;'5;LSTAT;CORRE'
1630 PRINT
1640 PRINT;;;'(EXPLAINED IN 5;LSTAT;REGEXP) - REQUESTS THE NUMBER OF'
1641 PRINT;;'VARIABLES AND IF A ZERO REGRESSION INTERCEPT IS TO BE'
1642 PRINT;;'FORCED.  IT ALSO ALLOWS THE USER TO LABEL EACH VARIABLE.'
1644 PRINT
1645 PRINT;;;'PRINTS THE MEANS, VARIANCES, STANDARD DEVIATIONS AND A'
1646 PRINT;;"CORRELATION MATRIX.  ENTERS THIS DATA INTO '$REG' AND "
1647 PRINT;;'LINKS TO 5;LSTAT;STP2.'
1650 PRINT
1660 PRINT;;'5;LSTAT;STP2'
1670 PRINT
1680 PRINT;;;'THIS PROGRAM PERFORMS THE ACTUAL STEPWISE REGRESSION.'
1690 PRINT;;'IT COMPUTES A SERIES OF REGRESSION EQUATIONS IN STEPS.'
1700 PRINT;;'AT EACH STEP, ONE VARIABLE IS ADDED OR DELETED, ACCORDING'
1710 PRINT;;'TO THE CONTROL DATA INPUT AT THE BEGINNING OF THIS PROGRAM
.'
1720 PRINT
1730 PRINT;;;'BEFORE ACTUAL COMPUTATION BEGINS, THE FOLLOWING CONTROL'
1740 PRINT;;'INFORMATION IS REQUESTED BY INPUT:'
1750 PRINT
1760 PRINT;;;;'F-RATIO FOR ACCEPTANCE OF A VARIABLE IN REGRESSION.'
1770 PRINT;;;;'F-RATIO FOR DELETION OF A VARIABLE FROM REGRESSION.'
1780 PRINT;;;'MINIMUM TOLERANCE FOR ACCEPTING A VARIABLE.'
1790 PRINT
1791 PRINT;;;'DEFAULT VALUES (OBTAINED BY ENTERING "0") FOR THE'
1792 PRINT;;;'ABOVE THREE VALUES ARE .01, .005 AND .001'
1793 PRINT;;;'RESPECTIVELY.'
1794 PRINT
1800 PRINT;;;'IN ADDITION, CONTROL VALUES FOR EACH VARIABLE MUST BE'
1810 PRINT;;'INPUT AS FOLLOWS:'
1820 PRINT
1830 PRINT;;;;'0 - DFLETE VARIABLE FROM ANALYSIS.'
1840 PRINT;;;;'1 - DEPENDENT VARIABLE - ONLY ONE ALLOWED'
1850 PRINT;;;;'2 - FREE VARIABLE - MAY BE USED IN THE ANALYSIS.'
1860 PRINT;;;;'3 TO 9 - FORCED VARIABLES - LOW TO HIGH LEVEL.'
1870 PRINT
1880 PRINT;;;'A FREE VARIABLE WILL BE ENTERED OR DELETED FROM REG-'
1890 PRINT;;'RESSION BASED ON THE PARTIAL F-RATIO AND TOLERANCE.  A'
1900 PRINT;;'HIGH LEVEL FORCED VARIABLE WILL BE ENTERED INTO REGRES-'
1910 PRINT;;'SION FIRST AND REMAIN IN REGRESSION.  LOWER LEVEL FORCED'
1920 PRINT;;'VARIABLES WILL ENTER REGRESSION ACCORDING TO LEVEL.'
1930 PRINT
1931 PRINT;;;'THE NUMBER OF STEPS WHICH MAY BE REQUIRED FOR COMPLETE'
1932 PRINT;;'ANALYSIS WILL THEN BE PRINTED.  THE USER MUST INPUT THE'
1933 PRINT;;'NUMBER OF STEPS HE WILL ALLOW WHICH SHOULD BE LESS THAN'
1934 PRINT;;'OR EQUAL TO THE PRINTED VALUE.'
1935 PRINT
```

```
1940 PRINT;;;'OUTPUT AT EACH STEP CONSISTS OF THE CONTROL INFORMATION,'
1950 PRINT;;'THE VARIABLE ENTERED OR DELETED, THE MULTIPLE CORRELATION'
1960 PRINT;;'COEFFICIENT, THE STANDARD ERROR OF THE ESTIMATE AND AN'
1970 PRINT;;'ANOVA TABLE.'
1980 PRINT
1990 PRINT;;;'FOR EACH VARIABLE IN REGRESSION, THE COEFFICIENT,'
2000 PRINT;;'STANDARD ERROR, T-VALUE, BETA COEFFICIENT, PARTIAL F-RATIO
,'
2010 PRINT;;'AND CONTROL VALUE WILL BE PRINTED.  IF ZERO REGRESSION '
2020 PRINT;;'INTERCEPT IS NOT REQUESTED, THEN THE INTERCEPT WILL'
2030 PRINT;;'ALSO BE PRINTED.'
2040 PRINT
2050 PRINT;;;'FOR EACH VARIABLE NOT IN REGRESSION, THE PARTIAL CORREL-'
2060 PRINT;;'LATION COEFFICIENT, TOLERANCE, PARTIAL F-RATIO AND CONTROL
'
2070 PRINT;;'VALUE WILL BE PRINTED.'
2080 PRINT
2090 PRINT;;;'UPON TERMINATION OF ENTERING AND DELETING VARIABLES, DEP-
'
2100 PRINT;;'ENDING ON THE CONTROL VALUES ENTERED, A SUMMARY TABLE WILL
'
2110 PRINT;;'BE PRINTED LISTING, BY STEP, THE VARIABLE ENTERED OR'
2120 PRINT;;'DELETED, THE MULTIPLE CORRELATION COEFFICIENT, STANDARD'
2130 PRINT;;'ERROR, F-RATIO AND NUMBER OF VARIABLES IN REGRESSION.'
2140 PRINT
2150 PRINT;;;'THE USER WILL THEN BE ASKED IF HE WISHES A RESIDUAL'
2160 PRINT;;'ANALYSIS.  IF "YES" IS ENTERED, STP2 WILL ENTER THE '
2170 PRINT;;"NECESSARY INFORMATION IN '$REG' AND LINK TO 5;LSTAT;"
2180 PRINT;;'RESID.  IF "NO" THEN THE PROGRAM WILL LINK TO 5;LSTAT;'
2190 PRINT;;'REGXFER.'
2200 PRINT
2210 PRINT;;'5;LSTAT;RESID'
2220 PRINT
2230 PRINT;;;'(EXPLAINED IN 5;LSTAT;REGEXP) - PROVIDES A LISTING OF'
2240 PRINT;;'ACTUAL, CALCULATED, RESIDUAL AND NORMAL DEVIATE RESIDUAL'
2250 PRINT;;'VALUES.  MUST BE RUN PRIOR TO 5;LSTAT;REGPLT.'
2260 PRINT
2270 PRINT;;'5;LSTAT;REGPLT'
2280 PRINT
2290 PRINT;;;'(EXPLAINED IN 5;LSTAT;REGEXP) - PROVIDES RESIDUAL AND'
2300 PRINT;;'SCATTER PLOTS FOR EACH VARIABLE IN REGRESSION, THE DEPEN-'
2310 PRINT;;'DENT VARIABLE AND CALCULATED VALUES.'
2320 PRINT
2330 PRINT;;'5;LSTAT;REGXFER'
2340 PRINT
2350 PRINT;;;'(EXPLAINED IN 5;LSTAT;REGEXP) - ALLOWS RUNNING '
2360 PRINT;;'ANOTHER PROBLEM USING THE SAME DATA FILE OR ENTERING'
2370 PRINT;;'ANOTHER DATA FILE FOR ANALYSIS, BOTH WITHOUT DELETING'
2380 PRINT;;"'$REG', OR FOR TERMINATING THE EXECUTION."
2390 PRINT
2400 PRINT;;;'NOW, GO TO THE EXEC AND COPY 5;LSTAT;REGEXP FOR A'
2410 PRINT;;'MORE COMPREHENSIVE EXPLANATION OF THE FOUR GENERAL'
2420 PRINT;;'PROGRAMS.'
-
```

```
COPY REGEXP TO TEL TEXT
```

********** 5;LSTAT;CORRE **********

THIS PROGRAM CALCULATES THE MEANS, VARIANCES AND STANDARD DEVIATIONS FOR EACH VARIABLE AND A CORRELATION MATRIX.

REQUESTED INPUT CONSISTS OF THE NUMBER OF OBSERVATIONS AND IF A ZERO REGRESSION INTERCEPT IS DESIRED (SUCH AS FOR SOLUTION OF SIMULTANEOUS EQUATIONS).

THE USER IS ALLOWED TO LABEL EACH VARIABLE WITH A CHARACTER SET OF NOT MORE THAN TEN CHARACTERS. THIS LABEL WILL BE USED TO IDENTIFY THE VARIABLES (IN ADDITION TO NUMBER) IN ALL SUCCEDING PROGRAMS.

IF A ZERO REGRESSION INTERCEPT IS REQUESTED, A WARNING THAT ALL VARIANCES, STANDARD DEVIATIONS AND CORRELATIONS ARE CALCULATED ABOUT THE ORIGIN RATHER THAN ABOUT THE MEAN IS PRINTED.

THE MEANS, VARIANCES AND STANDARD DEVIATIONS FOR EACH VARIABLE AND THE CORRELATION MATRIX IS THEN PRINTED AND THE PROGRAM LINKS TO THE PROPER REGRESSION ANALYSIS.

********** 5;LSTAT;RESID **********

THIS PROGRAM CALCULATES AND PRINTS THE ACTUAL AND CALCULATED VALUE OF THE DEPENDENT VARIABLE, THE RESIDUAL (ACTUAL - CALCULATED), AND THE RESIDUAL NORMAL DEVIATE FOR EACH OBSERVATION.

IT THEN PRINTS A HISTOGRAM OF THE RESIDUALS (NORMAL DEVIATES) AND TESTS THE UPPER AND LOWER EXTREME RESIDUALS BY THE DIXON CRITERION FOR OUTLIERS. IF THE NUMBER OF OBSERVATIONS IS 30 OR LESS, THE CRITICAL VALUE (ALPHA = .10) IS PRINTED.

THE THREE LOWEST RESIDUAL VALUES AND THE THREE HIGHEST RESIDUAL VALUES ARE THEN PRINTED ALONG WITH THE OBSERVATION NUMBER.

THE PROGRAM THEN REQUESTS A 'YES' OR 'NO' INPUT TO THE USER'S DESIRE FOR PLOTS. IF 'YES' IS INPUT, THE PROGRAM LINKS TO 5;LSTAT; REGPLT. IF 'NO' IS INPUT, THE PROGRAM LINKS TO 5;LSTAT;REGXFER.

********** 5;LSTAT;REGPLT *********

THIS PROGRAM PROVIDES 5" X 5" PLOTS IN THE FOLLOWING ORDER:

```
RESIDUALS (Y-AXIS) VS OBSERVATION NUMBER
    "              VS CALCULATED VALUES
    "              VS EACH VARIABLE IN REGRESSION
                      (INCL. DEPENDENT VARIABLE)
```

THE DEPENDENT VARIABLE THEN BECOMES THE Y-AXIS VARIABLE AND

THE CALCULATED VALUES REPLACE THE DEPENDENT VARIABLE IN SEQUENCE OF VARIABLES IN REGRESSION.

5" X 5" SCATTERPLOTS ARE THEN PRINTED FOR EACH VARIABLE IN REGRESSION AND THE CALCULATED VALUES ALONG THE X-AXIS VS THE DEPENDENT VARIABLE ALONG THE Y-AXIS.

THE PROGRAM THEN LINKS TO 5;LSTAT;REGXFER

********** 5;LSTAT;REGXFER **********

UPON ENTERING THIS PROGRAM, THE USER IS ASKED IF HE HAS ANOTHER PROBLEM USING THE SAME DATA FILE. IF 'YES' IS INPUT, 5;LSTAT;CORRE IS BYPASSED AND THE PROGRAM LINK DIRECTLY TO THE PROPER REGRESSION ANALYSIS PROGRAM.

IF 'NO' IS INPUT, THE USER IS ASKED TO INPUT A NEW DATA FILE NAME OR 'STOP' TO END THE SESSION. IF A NEW DATA FILE NAME IS INPUT, THE PROGRAM PERFORMS THE SAME STEPS AS THE INITIAL PROGRAM AND THEN LINKS TO 5;LSTAT;CORRE.

IF 'STOP' IS INPUT, A REMINDER TO DELETE $REG IS PRINTED AND THE PROGRAM ENDS.

```
QUIT
-COPY %CORRE TO TEL TEXT


1000 DEFAULT DOUBLE
1010 OPEN '$REG',2,BINARY,RANDOM,IO,OLD
1020 INPUT FROM 2 AT 1:HED$,F$,TYP
1030 OPEN F$,1,INPUT,OLD
1040 ON ENDFILE(1) GOTO 1390
1050 PRINT;'NUMBER OF VARIABLES':
1060 INPUT P
1070 DOUBLE A(P,P),B(P),S(P),U(P)
1080 STRING L$(P)
1090 PRINT;'FORCE ZERO REGRESSION INTERCEPT (YES OR NO)':
1100 INPUT AN$
1110 IF AN$='YES' THEN YINT=0 ELSE YINT=1
1120 PRINT
1130 PRINT;'LABELS<=10 CHAR. (ENTER NUMERIC "0" FOR BLANK):'
1140    FOR I=1 TO P
1150    PRINT;,'X(':I:') = ':
1160    INPUT L$(I)
1170    IF LENGTH(L$(I))<11 THEN 1190 ELSE PRINT;
          'TOO MANY CHARACTERS - TRY AGAIN'
1180    GOTO 1150
1190    IF L$(I)='0' THEN L$(I)=' '
1200    NEXT I
1210 PRINT IN FORM HED$:
1220 IF YINT THEN 1310
1230 MAT INPUT FROM 1:U
1240 N=N+1
1250    FOR I=1 TO P
1260    B(I)=B(I)+(U(I)-B(I))/N
1270       FOR J=1 TO I
1280       A(I,J)=A(I,J)+U(I)*U(J)
1290    NEXT J,I
1300 GOTO 1230
1310 MAT INPUT FROM 1:U
1320 N=N+1, NN=(N-1)/N
1330    FOR I=1 TO P
1340    U(I)=U(I)-B(I), B(I)=B(I)+U(I)/N
1350       FOR J=1 TO I
1360       A(I,J)=A(I,J)+NN*U(I)*U(J)
1370    NEXT J,I
1380 GOTO 1310
1390 PRINT IN FORM"6B'NUMBER OF OBSERVATIONS'6%/6B'NUMBER OF '
       'VARIABLES'9%/6B'FORCE ZERO INTERCEPT'8%///":N,P,AN$
1400 IF YINT THEN 1440 ELSE PRINT;;
      'WARNING...WHEN A ZERO REGRESSION INTERCEPT IS REQUESTED'

1410 PRINT,;' ALL VARIANCES, COVARIANCES, STANDARD DEVIATIONS,'

1420 PRINT,;' AND CORRELATIONS ARE COMPUTED ABOUT THE ORIGIN'
1430 PRINT,;' RATHER THAN ABOUT THE MEAN.'
1440 PRINT IN FORM"//6B
      'VARIABLE'11B'MEAN'8B'VARIANCE'8B'STD. DEV.'//":
1450 M=N-YINT, PROB=0
1460    FOR I=1 TO P
1470    SD2=A(I,I)/M, S(I)=SQRT(SD2), A(I,I)=1
1480    PRINT IN FORM'11% 3% 3(10%.5%)/':L$(I),I,B(I),SD2,S(I)
```

```
1490       FOR J=1 TO I-1
1500       A(J,I)=A(I,J)/(M*S(I)*S(J))
1510    NEXT J,I
1520 PRINT IN FORM"//6B'CORRELATION MATRIX'/":
1530    FOR I=1 TO P
1540    PRINT
1550    PRINT;;'ROW ':I
1560    PRINT;;
1570       FOR J=1 TO I
1580       PRINT IN FORM'4%.5%':A(J,I)
1590       IF NOT (J MOD 6) OR J=I THEN PRINT ELSE 1610
1600       IF J#I THEN PRINT;;
1610    NEXT J,I
1620 PRINT ON 2 AT 2:P,N,YINT,PROB
1630 MAT PRINT ON 2 AT 3:A
1640 MAT PRINT ON 2 AT 4:L$,B,S
1650 ON TYP GOTO 1660,1670
1660 LINK '5;LSTAT;MUL2'
1670 LINK '5;LSTAT;STP2'
```

COPY %STP2 TO TEL TEXT

```
1000 DEFAULT DOUBLE
1010 OPEN '$REG',1,IO,RANDOM,BINARY,OLD
1020 INPUT FROM 1 AT 1:HED$
1030 INPUT FROM 1 AT 2:P,N,YINT,PROB
1040 DOUBLE A(P,P),B(P),S(P),BETA(0:P),SUM(2*P,4),C(P),U(P)
1050 STRING L$(P)
1060 MAT INPUT FROM 1 AT 3:A
1070 MAT INPUT FROM 1 AT 4:L$,B,S
1080 PROB=PROB+1, M=N-YINT
1090 PRINT IN FORM"/%//3B'CONTROL DATA FOR PROBLEM NO.'3%//":
      CHAR(12),PROB
1100 PRINT;'F-LEVEL FOR INCLUSION (0=.01)':
1110 INPUT FIN
1120 IF FIN<=0 THEN FIN=.01
1130 PRINT;'F-LEVEL FOR DELETION (0=.005)':
1140 INPUT FOUT
1150 IF FOUT<=0 THEN FOUT=.005
1160 IF FIN<FOUT THEN PRINT;'F-IN < F-OUT - TRY AGAIN'
      ELSE 1180
1170 GOTO 1100
1180 PRINT;'TOLERANCE LEVEL (0=.001)':
1190 INPUT TOL
1200 IF TOL<=0 THEN TOL=.001
1210 PRINT
1220 PRINT;'VARIABLE CONROL VALUES'
1230 PRINT;;'0 - DELETE VARIABLE FROM ANALYSIS'
1240 PRINT;;'1 - DEPENDENT VARIABLE'
1250 PRINT;;'2 - FREE VARIABLE - MAY BE USED IN ANALYSIS'
1260 PRINT;;'3 TO 9 - FORCED VARIABLE - LOW TO HIGH LEVEL'
1270 PRINT
1280 D,Q,STP,CNT=0
1290 PRINT;'CONTROL VALUE FOR:'
1300    FOR I=1 TO P
1302       FOR J=1 TO I-1
1304       A(I,J)=A(J,I)
1306       NEXT J
1310    PRINT IN FORM"11% 3%' = '":L$(I),I
1320    INPUT C(I)
1330    IF C(I)>=0 AND C(I)<=9 THEN 1350 ELSE PRINT;;;
         'INVALID VALUE - TRY AGAIN'
1340    GOTO 1310
1350    ON C(I)+1 GOTO 1420, 1380
1360    CNT=CNT+1
1370    GOTO 1430
1380    IF D THEN PRINT;;;'ONLY ONE DEPENDENT VARIABLE ALLOWED':
         ' START OVER' ELSE 1400
1390    GOTO 1280
1400    D=I, SST=M*S(D)^2
1410    GOTO 1430
1420    C(I)=1
1430    NEXT I
1440 IF D THEN 1460 ELSE PRINT;
      'NO DEPENDENT VARIABLE SPECIFIED - TRY AGAIN'
1450 GOTO 1280
1460 IF CNT THEN 1480 ELSE PRINT;;;
```

```
     'NO INDEPENDENT VARIABLES SPECIFIED - TRY AGAIN'
1470 GOTO 1280
1480 PRINT
1490 PRINT;'THIS PROBLEM MAY REQUIRE UP TO ':2*CNT:
      ' STEPS TO SOLVE.'
1500 PRINT
1510 PRINT;'ENTER THE MAXIMUM NUMBER OF STEPS DESIRED FOR SOLUTION':
1520 INPUT DFLT
1530 STP=STP+1
1540 IF M=Q THEN 1610
       ELSE VOUT=(FOUT*A(D,D))/(M-Q)-7, VK=VOUT+9
1550    FOR I=1 TO P
1560    IF C(I)<1 THEN TEST=C(I)-A(I,D)^2/A(I,I) ELSE 1580
1570    IF TEST<VK THEN VK=TEST, K=I
1580    NEXT I
1590 IF VK<VOUT THEN FLAG=-1, C(K)=C(K)+9 ELSE 1610
1600 GOTO 1690
1610 VIN=(FIN*A(D,D))/(FIN+M-Q-1)+2, VK=VIN-9
1620    FOR I=1 TO P
1630    IF C(I)>1 AND A(I,I)>=TOL THEN TEST=C(I)+A(I,D)^2/A(I,I)
         ELSE 1650
1640    IF TEST>VK THEN VK=TEST, K=I
1650    NEXT I
1660 IF VK>=VIN AND M-Q-3+C(K)>0 THEN 1680 ELSE PRINT IN FORM"//6B
      'F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION.'/":

1670 GOTO 2080
1680 FLAG=1, C(K)=C(K)-9
1690 AKK=A(K,K), U(K)=-FLAG, A(K,K)=0
1700    FOR I=1 TO P
1710    IF I<K THEN U(I)=A(I,K), A(I,K)=0
1720    IF I>K THEN U(I)=A(K,I), A(K,I)=0
1730    NEXT I
1740    FOR J=1 TO P
1750       FOR I=1 TO J
1760       A(I,J),A(J,I)=A(I,J)-U(I)*U(J)/AKK
1770    NEXT I,J
1780 PRINT IN FORM HED$:
1790 SUM(STP,1)=K*FLAG, Q=Q+FLAG, DF=M-Q, SS=SST*A(D,D),
     RSS=SST-SS, RMS=RSS/Q
1800 IF DF THEN MS=SS/DF ELSE MS=0
1810 IF MS THEN SUM(STP,4),F=RMS/MS, SUM(STP,3),SE=SQRT(MS)
       ELSE SUM(STP,4),F,SUM(STP,3),SE=0
1820 SUM(STP,2),R=SQRT(1-A(D,D)), SUM(STP,3),SE=SQRT(MS),
     ALPH=B(D), A$,B$=' '
1830 IF L$(D)=A$ THEN 1850 ELSE A$=' ('+L$(D)+')'
1840 A$=A$+SPACE(18-LENGTH(A$))
1850 PRINT IN FORM"6B'PROBLEM NUMBER'10% 18B'F TO ENTER'9%.5%/
      6B'STEP NUMBER'13% 18B'F TO REMOVE'8%.5%/6B'DEPENDENT '
      'VARIABLE'6% 18%'TOLERANCE LEVEL'4%.5%//":PROB,FIN,
     STP,FOUT,D,A$,TOL
1860 IF FLAG=1 THEN F$='ENTERED' ELSE F$='REMOVED'
1870 IF L$(K)=B$ THEN 1890 ELSE B$=' ('+L$(K)+')'
1880 B$=B$+SPACE(15-LENGTH(B$))
1890 PRINT IN FORM"6B'VARIABLE'2(8%) 15%/6B'MULT. CORR. COEFF.'
      6%.5%/6B'STD. ERROR EST.'9%.5%///":F$,K,B$,R,SE
1900 PRINT IN FORM"34B'ANOVA TABLE'//23B'DF'6B'SUM OF SQ.'
      8B'MEAN SQ.'9B'F-RATIO'//6B'REGRESSION'9% 3(10%.5%)/6B
      'RESIDUAL'11% 2(10%.5%)/6B'TOTAL'14% 10%.5%///":Q,RSS,
      RMS,F,DF,SS,MS,M,SST
 1910 PRINT IN FORM"28B'VARIABLES IN REGRESSION'//6B'VARIABLE'
       4B'COEFFICIENT'8B'STD.  COMPUTED    BETA    F-OUT  TYP'
       /36B'ERROR   T- VALUE   COEFF.'//":
 1920    FOR I=1 TO P
 1930    IF C(I)>0 THEN 1950 ELSE BETA(I)=A(I,D)*S(D)/S(I)
 1940    IF YINT THEN ALPH=ALPH-BETA(I)*B(I)
 1950    NEXT I
 1960 IF YINT THEN PRINT IN FORM"5B'INTERCEPT'9%.5%/":ALPH
 1970    FOR I=1 TO P
 1980    IF C(I)>0 THEN 2010 ELSE VS=SE/S(I)*SQRT(ABS(A(I,I))/M)
 1990    IF VS THEN CTV=BETA(I)/VS, FRAT=CTV^2
          ELSE CTV,FRAT=0
 2000    PRINT IN FORM"11% 3% 9%.5% 6%.5% B 2(5%.3%)2B #.6# P70
          '('%')'/":L$(I),I,BETA(I),VS,CTV,BETA(I)*S(I)/S(D),
          FRAT,C(I)+9
 2010    NEXT I
 2020 IF Q=CNT THEN 2060 ELSE PRINT IN FORM"//23B'VARIABLES NOT IN '
       'REGRESSION'//6B'VARIABLE'6B'PART. CORR.  TOLERANCE'   F-IN'
       3B'TYP'//":
 2030    FOR I=1 TO P
 2040    IF C(I)>1 THEN PRINT IN FORM"11% 3% 10%.5% 8%.3% 3B
          #.6# P53'('%')'/":L$(I),I,A(I,D)/SQRT(A(I,I)*A(D,D)),
          A(I,I),A(I,D)^2*(M-Q-1)/(A(I,I)*A(D,D)-A(I,D)^2),C(I)
 2050    NEXT I
 2060 IF STP<DFLT THEN 1530 ELSE PRINT IN FORM"//6B
       'DEFAULT VALUE REACHED - COMPUTATION TERMINATED.'/":
 2070 STP=STP+1
 2080 PRINT IN FORM HED$:
 2090 PRINT IN FORM"33B'SUMMARY TABLE'//7B'STEP'8B'VARIABLE'
       6B'MULT.'8B'STD.'10B'F'6B'NO. IN'/6B'NUMBER'5B'NAME'
       '  IN OUT    CORR.'7B'ERROR'8B'RATIO     REG.'//":
 2100 TOT=0
 2110    FOR I=1 TO STP-1
 2120    NS=SUM(I,1), TOT=TOT+SGN(NS)
 2130    IF NS<0 THEN A$=' ', NS=-NS, B$=STR(NS)
          ELSE A$=STR(NS), B$=' '
 2140    PRINT IN FORM'10% 11% 2(4%) 3%.5% 2(7%.5%) 7%/':
          I,L$(NS),A$,B$,SUM(I,2),SUM(I,3),SUM(I,4),TOT
 2150    NEXT I
 2155 PRINT ON 1 AT 2:P,N,YINT,PROB
 2160 PRINT IN FORM"/////6B'RESIDUAL ANALYSIS (YES OR NO)'":
 2170 INPUT AN$
 2180 IF AN$='NO' THEN LINK '5!LSTAT!REGXFER'
 2190    FOR I=1 TO P
 2200    IF C(I)>0 THEN BETA(I)=0
 2210    NEXT I
 2220 IF YINT THEN BETA(0)=ALPH ELSE BETA(0)=0
 2230 TOT=-INT(-TOT/3)
 2250 PRINT ON 1 AT 5:D,MS,TOT
 2260 MAT PRINT ON 1 AT 6:BETA
 2270 LINK '5!LSTAT!RESID'
-
```

```
COPY %RESID TO TEL TEXT


1000 DEFAULT DOUBLE
1010 OPEN '$REG',2,IO,RANDOM,BINARY,INPUT,OLD

1020 INPUT FROM 2 AT 1:HED$,F$,TYP
1030 INPUT FROM 2 AT 2:P,N,YINT,PROB
1040 INPUT FROM 2 AT 5:D,MS,TOT
1050 DOUBLE B(0:P),R(N),A(2,3),NA(2,3),MM(P,2),Q(P),RD(30)
1060 STRING L$(P),H$(TOT)
1070 MAT INPUT FROM 2 AT 4:L$
1080 MAT INPUT FROM 2 AT 6:B
1090    FOR I=1 TO P
1100    IF B(I) OR I=D THEN A$=' '+STR(I) ELSE 1130
1110    IF I<10 THEN A$=' '+A$
1120    IF L$(I)=' ' THEN L$(I)=A$+SPACE(11)
         ELSE L$(I)=A$+'-'+L$(I)+SPACE(10-LENGTH(L$(I)))
1130    NEXT I
1140 IF PROB<10 THEN TIT$=STR(PROB)+' '
      ELSE TIT$=STR(PROB)
1150 TIT$="6B'PROBLEM NUMBER "+TIT$+"'17B'DEPENDENT VARIABLE"
      +L$(D)+"'//"+STR(TOT)+"(72%/)/"
1160 H$(1)='INDEPEND. VARIABLES ', FF=0, J=1
1170    FOR I=1 TO P
1180    IF B(I)=0 THEN 1220
1190    IF (FF MOD 3) THEN H$(J)=H$(J)+', '
1200    H$(J)=H$(J)+L$(I), FF=FF+1
1210    IF LENGTH(H$(J))>60 AND J<TOT
         THEN J=J+1, H$(J)=SPACE(20)
1220    NEXT I
1230 IF LENGTH(H$(J))=20 THEN J=J-1
      ELSE H$(J)=H$(J)+SPACE(66-LENGTH(H$(J)))
1240 GOSUB 2010
1250 OPEN F$,1,INPUT,OLD
1260 ND=1/SQRT(MS)
1270    FOR I=1 TO 3
1280    A(1,I)=1E11, A(2,I)=-1E11
1290    NEXT I
1300    FOR I=1 TO P
1310    MM(I,1)=-1E11, MM(I,2)=1E11
1320    NEXT I
1330 CMAX=-1E11, CMIN=1E11
1340    FOR I=1 TO N
1350    MAT INPUT FROM 1:Q
1360    R(I)=Q(D)-B(0), L,U=0
1370       FOR J=1 TO P
1380       MM(J,1)=MAX(MM(J,1),Q(J)), MM(J,2)=MIN(MM(J,2),Q(J))

1390       IF B(J) THEN R(I)=R(I)-B(J)*Q(J)
1400       NEXT J
1410    PRINT IN FORM'11% 4(8%.5%)/':I,Q(D),Q(D)-R(I),R(I),R(I)*ND
1420    IF NOT (I MOD 10) THEN PRINT
1430    CMAX=MAX(CMAX,Q(D)-R(I)), CMIN=MIN(CMIN,Q(D)-R(I))
1440    IF (I MOD 40) THEN 1460
1450 GOSUB 2010
1460    IF R(I)>A(1,1) THEN 1490
         ELSE A(1,1)=R(I), NA(1,1)=I
1470    IF R(I)>A(1,2) THEN 1490
```

```
          ELSE A(1,1)=A(1,2), NA(1,1)=NA(1,2),
                A(1,2)=R(I)   , NA(1,2)=I
1480     IF R(I)<=A(1,3) THEN A(1,2)=A(1,3), NA(1,2)=NA(1,3),
                              A(1,3)=R(I)   , NA(1,3)=I
1490     IF R(I)<A(2,1) THEN 1520
          ELSE A(2,1)=R(I), NA(2,1)=I
1500     IF R(I)<A(2,2) THEN 1520
          ELSE A(2,1)=A(2,2), NA(2,1)=NA(2,2),
                A(2,2)=R(I)   , NA(2,2)=I
1510     IF R(I)>=A(2,3) THEN A(2,2)=A(2,3), NA(2,2)=NA(2,3),
                              A(2,3)=R(I)   , NA(2,3)=I
1520     NEXT I
1530 FF=0
1540 GOSUB 2010
1550 PRINT
1560 PRINT,'HISTOGRAM OF RESIDUALS (NORMAL DEVIATES)'
1570 PRINT
1580 K1=ROUND(4*A(1,3)*ND), K2=ROUND(4*A(2,3)*ND)
1590 INTEGER PLOT(K1:K2)
1600     FOR I=1 TO N
1610     I1=ROUND(4*R(I)*ND), PLOT(I1)=PLOT(I1)+1
1620     NEXT I
1622 IMAX=0
1623    FOR I=K1 TO K2
1624    IMAX=MAX(IMAX,PLOT(I))
1625    NEXT I
1626 IF IMAX<51 THEN 1630 ELSE K3=-INT(-50/IMAX)
1627 MAT PLOT=(K3)*PLOT
1630     FOR I=K1 TO K2
1640     IF (I MOD 4) THEN I$=':' ELSE I$=STR(.25*I)+' +'
1650     PRINT IN FORM'12%':I$
1660        FOR J=1 TO ROUND(PLOT(I))
1670        PRINT '*':
1680        NEXT J
1690     PRINT
1700     NEXT I
1710 DATA 000, 000,.886,.679,.557,.482,.434,.479,.441,.409,
        .517,.490,.467,.492,.472,.454,.438,.424,.412,.401,
        .391,.382,.374,.367,.360,.354,.348,.342,.337,.332
1720 MAT READ RD
1730 PRINT IN FORM"//14B'TEST FOR EXTREME RESIDUAL VALUES '
      ' (DIXON CRITERION)'//":
1740 IF N<31 THEN PRINT IN FORM"19B
      'CRITICAL VALUE (ALPHA = .10) ='%.3%' (MAX)'//":RD(N)
1750 IF N<14 THEN 1790 ELSE $X='R22'
1760 RS=(A(1,1)-A(1,3))/(A(2,1)-A(1,3))
1770 RL=(A(2,3)-A(2,1))/(A(2,3)-A(1,1))
1780 GOTO 1900
1790 IF N<11 THEN 1830 ELSE $X='R21'
1800 RS=(A(1,1)-A(1,3))/(A(2,2)-A(1,3))
1810 RL=(A(2,3)-A(2,1))/(A(2,3)-A(1,2))
1820 GOTO 1900
1830 IF N<8 THEN 1870 ELSE $X='R11'
1840 RS=(A(1,2)-A(1,3))/(A(2,2)-A(1,3))
1850 RL=(A(2,3)-A(2,2))/(A(2,3)-A(1,2))
1860 GOTO 1900
1870 $X='R10'
1880 RS=(A(1,2)-A(1,3))/(A(2,3)-A(1,3))
1890 RL=(A(2,3)-A(2,2))/(A(2,3)-A(1,3))
1900 PRINT IN FORM"19B 3%' (MIN) ='%.3% 12%' (MAX) ='%.3%//":
     $X,RS,$X,RL
1910   FOR I=0 TO 2
1920   X$1='R('+STR(NA(1,3-I))+')  =',
       X$2='R('+STR(NA(2,3-I))+')  ='
1930   PRINT IN FORM"25% 6%.3% 16% 6%.3%/":
        X$1,A(1,3-I),X$2,A(2,3-I)
1940   NEXT I
1950 PRINT IN FORM"////6B'DO YOU WISH PLOTS (YES OR NO)'":
1960 INPUT AN$
1970 IF AN$='NO' THEN LINK '5;LSTAT;REGXFER'
1980 PRINT ON 2 AT 7:A(2,3),A(1,3),CMAX,CMIN,TOT,TIT$
1990 MAT PRINT ON 2 AT 8:MM,R,H$
2000 LINK '5;LSTAT;REGPLT'
2010 PRINT IN FORM HED$:
2020 MAT PRINT IN FORM TIT$:H$
2030 IF FF THEN PRINT IN FORM"2(12%) 9B'CALC.'7B'RESIDUAL'6B
      'NORM.DEV'//":'OBS.','ACT.'
2040 RETURN
-
```

```
COPY %REGPLT TO TEL TEXT

1000 DEFAULT DOUBLE !

        ********** INPUT DATA **********

1010 OPEN '$REG',1,IO,RANDOM,BINARY,OLD
1020 INPUT FROM 1 AT 1:HED$,F$,TYP
1030 INPUT FROM 1 AT 2:P,N,YINT,PROB
1040 INPUT FROM 1 AT 5:D
1050 INPUT FROM 1 AT 7:RMAX,RMIN,CMAX,CMIN,H1,TIT$
1060 DOUBLE X(N,P),R(N),Y(N),MM(P,2),B(0:P),QQ(0:5)
1070 INTEGER A(0:30,0:50)
1080 STRING L$(P),LBL(0:30),V$(10),H$(H1),Q(0:50)
1090 MAT INPUT FROM 1 AT 4:L$
1100 MAT INPUT FROM 1 AT 6:B
1110 MAT INPUT FROM 1 AT 8:MM,R,H$
1120 OPEN F$,2,INPUT,OLD
1130 MAT INPUT FROM 2:X !

        ********** SCALING FUNCTION **********

1140 DEF DOUBLE FNS(LL,UL,KK)
1150 DOUBLE LL,UL,KK
1160 DEFAULT DOUBLE
1170 RNG=UL-LL, P10=10^(INT(LOG10(RNG))), RNG=RNG/P10
1180 SL=SGN(LL), SU=SGN(UL), LL=LL/P10, UL=UL/P10, TR=0
1190 IF SL+SU<0 THEN UL=-LL, LL=UL-RNG
1200 IF UL<=INT(LL)+10 THEN 1220
      ELSE P10=10*P10, LL=.1*LL, UL=.1*UL
1210 GOTO 1200
1220 TL=INT(LL), TR=10, TL1=INT(2*LL/3)*1.5
1230 IF UL>TL1+7.5 THEN 1320
      ELSE TL=TL1, TR=7.5, TL1=INT(2*LL)*.5
1240 IF UL>TL1+5 THEN 1320
      ELSE TL=TL1, TR=5, TL1=INT(10*LL/4)*.4
1250 IF UL>TL1+4 THEN 1320
      ELSE TL=TL1, TR=4, TL1=INT(10*LL/3)*.3
1260 IF UL>TL1+3 THEN 1320
      ELSE TL=TL1, TR=3, TL1=INT(4*LL)/4
1270 IF UL>TL1+2.5 THEN 1320
      ELSE TL=TL1, TR=2.5, TL1=INT(5*LL)*.2
1280 IF UL>TL1+2 THEN 1320
      ELSE TL=TL1, TR=2, TL1=INT(20*LL/3)*.15
1290 IF UL>TL1+1.5 THEN 1320
      ELSE TL=TL1, TR=1.5, TL1=INT(4*LL)/4
1300 IF UL>TL1+1.25 THEN 1320
      ELSE TL=TL1, TR=1.25, TL1=INT(10*LL)*.1
1310 IF UL<=TL1+1 THEN TL=TL1, TR=1
1320 IF UL+LL=0 THEN TL=-TR/2
1330 IF SL+SU=0 THEN 1360
1340 IF UL<=TR THEN TL=0
1350 IF SL+SU<0 THEN TL=-TL-TR
1360 LL=TL*P10, RNG=TR*P10
1370 IF KK=30 THEN SY=30/(TR*P10), Y0=ROUND(-LL*SY)
      ELSE SX=50/(TR*P10), X0=ROUND(-LL*SX)
1380 RETURN LL
1390 END !
```

```
            ********** ASSIGN CONSTANTS **********

1400     FOR I=2 TO 9
1410     V$(I)=STR(I)
1420     NEXT I
1430 V$(1)='*', V$(10)='0', LBL(0)=' ',
     FMT$='7% 6%.%% 52%/', TIT$=TIT$+'///'
1431     FOR I=1 TO P
1432     IF B(I) OR I=D THEN A$='VARIABLE '+STR(I)
          ELSE 1434
1433     IF L$(I)=' ' THEN L$(I)=A$
          ELSE L$(I)=A$+' ('+L$(I)+')'
1434     NEXT I
1440     FOR I=1 TO 10
1450     LBL(I),LBL(I+10),LBL(I+20)=' '
1460     NEXT I !

          ********** RESIDUAL PLOTS **********

1470 DATA R,E,S,I,D,U,A,L,S,
1480     FOR I=19 TO 11 STEP -1
1490     READ LBL(I)
1500     NEXT I
1510 S$=' VS RESIDUALS', UL=MAX(RMAX,-RMIN),
     YL=FNS(-UL,UL,30)
1520     FOR I=1 TO N
1530     Y(I)=ROUND((R(I)-YL)*SY)
1540     NEXT I !

          ********** TIME - RESIDUAL **********

1550 MAT A=(0)
1560 B$='OBSERVATION', A$=B$+'S'+S$, XL=FNS(0,N,50)
1570     FOR I=1 TO N
1580     XX=ROUND(I*SX), A(Y(I),XX)=A(Y(I),XX)+1
1590     NEXT I
1600 GOSUB 2050 !

          ********** CALCULATED - RESIDUAL **********

1610 MAT A=(0)
1620 B$='CALCULATED', A$=B$+S$, XL=FNS(CMIN,CMAX,50)
1630     FOR I=1 TO N
1640     XX=ROUND((X(I,D)-R(I)-XL)*SX)
1650     A(Y(I),XX)=A(Y(I),XX)+1
1660     NEXT I
1670 GOSUB 2050 !

          ********** ACTUAL - RESIDUAL **********

1680     FOR I=1 TO P
1690     IF B(I)=0 AND I#D THEN 1770
          ELSE XL=FNS(MM(I,2),MM(I,1),50)
1700     B$=L$(I), A$=B$+S$
1710     MAT A=(0)
1720        FOR J=1 TO N
1730        XX=ROUND((X(J,I)-XL)*SX)
1740        A(Y(J),XX)=A(Y(J),XX)+1
1750        NEXT J
1760     GOSUB 2050
```

```
1770     NEXT I !

           ********** INDEPENDENT - DEPENDENT **********

1780 DATA D,E,P,E,N,D,E,N,T,' ',' ',V,A,R,I,A,B,L,E,
1790     FOR I=24 TO 6 STEP -1
1800     READ LBL(I)
1810     NEXT I
1820 S$=' VS '+L$(D), YL=FNS(MM(D,2),MM(D,1),30),
     L$(D)='CALCULATED', MM(D,2)=CMIN, MM(D,1)=CMAX
1830     FOR I=1 TO N
1840     Y(I)=ROUND((X(I,D)-YL)*SY), X(I,D)=X(I,D)-R(I)
1850     NEXT I
1860 B(D)=1
1870     FOR I=1 TO P
1880     IF B(I)=0 THEN 2000
1910     B$=L$(I), A$=B$+S$, XL=FNS(MM(I,2),MM(I,1),50)
1940     MAT A=(0)
1950        FOR J=1 TO N
1960        XX=ROUND((X(J,I)-XL)*SX)
1970        A(Y(J),XX)=A(Y(J),XX)+1
1980        NEXT J
1990     GOSUB 2050
2000     NEXT I
2010 PRINT CHAR(12)
2020 PRINT
2030 LINK '5;LSTAT;REGXFER'
2040 END
2050 PRINT IN FORM HED$:
2060 MAT PRINT IN FORM TIT$:H$
2070 PRINT TAB(43-LENGTH(A$)/2):A$
2080 PRINT
2090     FOR II=30 TO 0 STEP -1
2100     IF (II MOD 6) THEN AE$=':' ELSE AE$='+'
2110     IF (II MOD 30) AND Y0#II THEN AL$,AX$=' '
          ELSE AL$='-', AX$='+'
2120        FOR JJ=0 TO 50
2130        IF A(II,JJ) THEN Q(JJ)=V$(MIN(A(II,JJ),10))
             ELSE 2150
2140        GOTO 2170
2150        IF (JJ MOD 10) THEN Q(JJ)=AL$ ELSE Q(JJ)=AX$
2160        IF JJ=0 OR JJ=50 OR JJ=X0 THEN Q(JJ)=AE$
2170        NEXT JJ
2180     IF (II MOD 6) THEN PRINT IN FORM '7% 10B':LBL(II)
          ELSE PRINT IN FORM'7% 5%.3% B':LBL(II),YL+II/SY
2185     MAT PRINT IN FORM'51(%)/':Q
2190     NEXT II
2200 PRINT
2210 PRINT TAB(10):
2220     FOR II=0 TO 5
2230     PRINT IN FORM'6%.3%':XL+10*II/SX
2240     NEXT II
2250 PRINT
2260 PRINT
2270 PRINT TAB(43-LENGTH(B$)/2):B$
2280 PRINT
2290 RETURN
-
```

```
COPY %REGXFER TO TEL TEXT


1000 OPEN '$REG',1,IO,BINARY,RANDOM,OLD
1010 DOUBLE TYP
1020 INPUT FROM 1 AT 1:HED$,F$,TYP
1030 PRINT;'ANOTHER PROBLEM USING THE SAME DATA (YES OR NO)':
1040 INPUT AN$
1050 IF AN$='NO' THEN 1170
1055 IF TREC(1)<5 THEN 1070
1060 ERASE FILE 1 FROM 5 TO TREC(1)
1070 ON TYP GOTO 1100, 1120
1080 REM
1090 REM
1100 REM
1110 REM
1120 LINK '5;LSTAT;STP2'
1130 REM
1140 REM
1150 REM
1160 REM
1170 PRINT;"INPUT NEW DATA FILENAME OR 'STOP'":
1180 INPUT F$
1190 ERASE FILE 1 FROM 1 TO TREC(1)
1200 IF F$='STOP' THEN 1390
1210 PRINT;'NEW JOB TITLE':
1220 INPUT HED$
1230 ON TYP GOTO 1240, 1280
1240 REM
1250 REM
1260 REM
1270 REM
1280 HED$="'"+CHAR(12)+"'//6B'STEPWISE MULTIPLE LINEAR REGRESSION'"
      +"///6B'"+HED$+"'///"
1290 GOTO 1340
1300 REM
1310 REM
1320 REM
1330 REM
1340 PRINT ON 1 AT 1:HED$,F$,TYP
1350 LINK '5;LSTAT;CORRE'
1360 REM
1370 REM
1380 REM
1390 PRINT
1400 PRINT;;"GOOD-BYE.  BE SURE TO DELETE FILE '$REG' BEFORE LEAVING"
1410 PRINT;'THE SYSTEM.'
_
```

5;LAB;MATHISTO

```
1000 PRINT;"DATA FILENAME OR 'EXP' FOR PROGRAM INFORMATION":
1010 INPUT F$ !

        ***** EXPLAIN PROGRAM *****

1020 IF F$#'EXP' THEN 1100
1030 OPEN '5;LSTAT;STATEXP',1,RANDOM,INPUT,OLD
1040    FOR I=160 TO 206
1050    INPUT FROM 1 AT I:A$
1060    IF A$=' ' THEN PRINT ELSE PRINT;A$
1070    NEXT I
1080 CLOSE 1
1090 END !

         ***** INITIALIZE - READ DATA *****

1100 DIM X(1001),W(12),$TAT(12),Y(100)
1110 MAT READ $TAT
1120 DATA OBSERVATIONS,MEAN     ,STD.DEV.,
          MINIMUM      ,RANGE    ,MAXIMUM ,
          VARIANCE     ,SKEWNESS,KURTOSIS,
          COEFF. VAR.  ,AVG.DEV.,RMS DEV.,
1130 OPEN F$,1,INPUT,OLD
1135 DEFAULT DOUBLE
1140 ON ENDFILE(1) GOTO 2320
1150 PRINT;'JOB TITLE - NO. ROWS -NO. COLUMNS':
1160 INPUT HED$,NROW,NCOL
1162 DIM XX(NROW,NCOL)
1164 PRINT;'LOG TRANSFORM?  INPUT  1 FOR YES  2 FOR NO ':
1166 INPUT ANS
1170 MAT INPUT FROM 1:XX
1180 FOR JJ=1 TO NCOL
1181 LL=1E11, UL=-1E11, XBAR,U2,U3,U4,W(11)=0, N=1
1182 FOR II=1 TO NROW
1183 IF XX(II,JJ)<0 THEN 1220
1184 IF ANS=2 THEN X(N)=XX(II,JJ) ELSE 1190
1186 GOTO 1200
1190 IF XX(II,JJ)<=0 THEN 1220 ELSE X(N)=LOG10(XX(II,JJ))
1200 XBAR=XBAR+(X(N)-XBAR)/N
1210 LL=MIN(X(N),LL), UL=MAX(X(N),UL)
1212 N=N+1
1220 NEXT II
1221 IF N=1 THEN 1228
1225 GOSUB 1240
1228 NEXT JJ
1230 CLOSE 1
1235 DO 1310,2330
1240    FOR I=1 TO N-1
1250    X1=X(I)-XBAR, X2=X1*X1
1260    U2=U2+X2, U3=U3+X1*X2, U4=U4+X2*X2
1270    W(11)=W(11)+ABS(X1)
1280    NEXT I
1290 W(1),N=N-1, W(2)=XBAR, W(4)=LL, W(5),RNG=UL-LL, W(6)=UL,
     W(7)=U2/(N-1), W(3)=SQRT(W(7)), W(8)=SQRT(N)*U3/(U2^1.5),
     W(9)=N*U4/(U2*U2), W(10)=100*W(3)/XBAR, W(11)=W(11)/N,
     W(12),RMS=SQRT(U2/N)

1300 REM !
        ***** OUTPUT PARAMETERS *****
```

```
1310 PRINT CHAR(12)
1320 PRINT
1330 PRINT;'SAMPLE STATISTICS'
1340 PRINT
1350 PRINT ;HED$;"PROBE ":JJ
1360 PRINT
1370 PRINT
1380    FOR I=1 TO 12
1390    PRINT;$TAT(I):'=':
1400    IF ABS(W(I))<10 AND ABS(W(I))>=1
         THEN PRINT IN FORM"3%.3%'E+00'":W(I)
         ELSE PRINT IN FORM'B #.8#':W(I)
1410    IF NOT (I MOD 3) THEN PRINT IN FORM'/':ELSE PRINT' ':
1420    NEXT I !

          ***** CALCULATE INITIAL HSTOGRAM *****

1430 NCEL=ROUND(1+10*LOG10(N)/3), RFAC=10^(-INT(LOG10(RNG)))
1440 TLL=INT(LL*RFAC)/RFAC, TINT=(UL-TLL)*RFAC/NCEL
1450 IF TINT>1 AND TINT<3 THEN TINT=-INT(-2*TINT)/2
      ELSE TINT=-INT(-TINT)
1460 IF TINT=7 OR TINT=9 THEN TINT=(TINT+1)/RFAC
       ELSE TINT=TINT/RFAC
1470 IF RNG>.8*NCEL*TINT THEN 1530 ELSE TINT=TINT*RFAC
1480 IF TINT<1.5 THEN TINT=10*TINT, FLAG=1 ELSE FLAG=0
1490 IF TINT<3 THEN TINT=TINT-.5 ELSE TINT=TINT-1
1500 IF TINT>6 THEN TINT=TINT-1
1510 IF FLAG THEN TINT=TINT/(10*RFAC) ELSE TINT=TINT/RFAC
1520 GOTO 1470 !

1530 TLL=TLL-.5*TINT
1540 IF LL>TLL+TINT THEN TLL=TLL+TINT ELSE 1570
1550 GOTO 1540
1570 NCEL=-INT((TLL-UL)/TINT)
1580 MAT Y=(0)
1590    FOR I=1 TO N
1600    Y0=(X(I)-TLL)/TINT+1, Y1=FIX(Y0)
1610    IF Y1<Y0 THEN Y(Y1)=Y(Y1)+1
         ELSE Y(Y1)=Y(Y1)+.5, Y(Y1-1)=Y(Y1-1)+.5
1620 NEXT I !

          ***** OUTPUT HISTOGRAM *****

1622 IF ANS=2 THEN L$=' ' ELSE L$=' LOG '
1630 PRINT IN FORM'////':
1640 PRINT,NCEL:' CELLS -':L$:'CELL INTERVAL = ':TINT
1650 PRINT IN FORM"/4B'MIDPOINT'3B'NO. OBS.'5B'% TOTAL'5B'TOT.CUM.'
      4B'Z-SCORE(RMS)'//":
1660 ZINT=TINT/RMS, CMP=TLL+.5*TINT, ZSC=(CMP-XBAR)/RMS, PP,CUM=0
1670    FOR I=1 TO NCEL
1675 TOT=100*Y(I)/(N+1),CUM=CUM+TOT
1677 IF ANS=2 THEN CPP=CMP ELSE CPP=10^CMP
1680    PRINT IN FORM'3B #.8# P15 6% 3(9%.3%)//':
         CPP,Y(I),TOT,CUM,ZSC
1690    CMP=CMP+TINT, ZSC=ZSC+ZINT, PP=MAX(Y(I),PP)
1700    NEXT I !

          ***** WHAT NEXT? *****
```

```
1705 GOTO 2325
1710 PRINT IN FORM"/////3B'ENTER - 1 FOR PLOT'/11B
      '2 FOR NEW CELL PARAMETERS'/11B'3 FOR NEXT FILE'
      /11B'4 FOR STOP'//11B'WHICH'":
1720 INPUT ANS
1730 ON ANS GOTO 1840, 1780, 1750
1740 END !

          ***** NEW DATA FILE *****

1750 PRINT;'NEW DATA FILENAME':
1760 INPUT F$
1770 GOTO 1130 !

          ***** NEW HSTOGRAM PARAMETERS *****

1780 PRINT;'NEW CELL MIDPOINT,INTERVAL':
1790 INPUT CMP,TINT
1800 DO 1310: 1350
1810 TLL=CMP-.5*TINT
1820 IF LL<TLL THEN TLL=TLL-TINT ELSE 1540
1830 GOTO 1820 !

          ***** PLOT HISTOGRAM *****

1840 NN=N, Z0=TINT/RMS*NN/SQRT(2*PI), PP=MAX(Z0,PP), PWR=1
1850 PWR=PWR-1, IND=PP*10^PWR
1860 IF IND<5.2 THEN PP=10^PWR, IND=1 ELSE 1880
1870 GOTO 2030
1880 IF IND<10.4 THEN PP=.5*10^PWR, IND=2 ELSE 1900
1890 GOTO 2030
1900 IF IND<26 THEN PP=.2*10^PWR, IND=3 ELSE 1850
1910 GOTO 2030 !

1920 PRINT IN FORM'4% B 4%':A,B
1930 IF FLG THEN PRINT'^' ELSE PRINT
1940 FLG=0
1950 RETURN

1960     FOR B=U1 TO U2 STEP U3
1970     A,FLG=X1
1980     GOSUB 1920
1990     A=A+U4
2000     GOSUB 1920
2010     NEXT B
2020 RETURN !

2030 NN=N, Z0=TINT/RMS*NN/SQRT(2*PI), PP=MAX(Z0,PP), PWR=1
2040 PWR=PWR-1, IND=PP*10^PWR
2050 IF IND<5.2 THEN PP=10^PWR, IND=1 ELSE 2070
2060 GOTO 2100
2070 IF IND<10.4 THEN PP=.5*10^PWR, IND=2 ELSE 2090
2080 GOTO 2100
2090 IF IND<26 THEN PP=.2*10^PWR, IND=3 ELSE 2040
2100 PRINT
2110 PRINT;'VERTICAL SCALE = ':ROUND(1/PP):' OBSERVATIONS/INCH'
2120 PRINT 'PLTL'
2130 NC=NCEL, CW=8000/NC, A=750, B,YY=9999/7, Z0=Z0*PP
2140 GOSUB 1920
2150 A=A+250-CW
```

```
2160     FOR I=1 TO NC
2170     A=A+2*CW
2180     IF Y(I) THEN 2200 ELSE A=A-CW
2190     GOTO 2270
2200     GOSUB 1920
2210     B=YY*(1+PP*Y(I))
2220     GOSUB 1920
2230     A=A-CW
2240     GOSUB 1920
2250     B=YY
2260     GOSUB 1920
2270     NEXT I
2280 A=A+250+CW
2290 GOSUB 1920
2300 PRINT,'PLTT'
2320 DO 1310
2325 GOTO 1228
2330 END
```

5;LSTAT;SUMMARY

```
100 PRINT;;'***** MODIFIED SEPTEMBER, 1974 *****'
110 PRINT
1000 PRINT
1010 PRINT;;"DATA FILENAME OR 'EXP' FOR PROGRAM DETAILS":
1020 INPUT F$
1030 PRINT
1040 IF F$='EXP' THEN 1960
1050 OPEN F$,1,INPUT,OLD
1060 ON ENDFILE(1) GOTO 1810
1070 PRINT;;"NUMBER OF APCD'S, NUMBER OF SITES":
1080 INPUT APCD,SITE
1090 NS=APCD+SITE
1100 PRINT;;'NUMBER OF DAYS':
1110 INPUT ND
1120 PRINT;;'NUMBER OF READINGS/DAY AT EACH SITE':
1130 INPUT NR
1140 DIM X(ND,NR,NS),H$(2,NS),T$(6),SRT(ND),Y(4,NS),CV(ND),S$(NS)
1150 MAT INPUT FROM 1:X
1160 PRINT;;'EARLIEST READING (MILITARY TIME)':
1170 INPUT TIME
1180 PRINT
1190 PRINT;;'SITE ID, MINUTES AFTER HOUR READ (2 CHAR EACH)'
1200     FOR I=1 TO NS
1210     IF I>APCD THEN S$(I)='  SITE' ELSE S$(I)='  APCD'
1220     PRINT;;S$(I):' ':
1230     IF I>APCD THEN PRINT I-APCD: ELSE PRINT I:
1240     INPUT H$(1,I),H$(2,I)
1250     IF LENGTH(H$(1,I))=2 AND LENGTH(H$(2,I))=2 THEN 1270            EL
SE PRINT;;'2 CHARACTERS FOR EACH - TRY AGAIN'
1260     GOTO 1220
1270     H$(1,I)='    '+H$(1,I)
1275     NEXT I
1280 PRINT
1290 PRINT;;'INPUT 6 LINES FOR HEADING (0=BLANK LINE)'
1300     FOR I=1 TO 6
1310     PRINT;;'LINE ':I:
1320     INPUT T$(I)
1330     IF T$(I)#'0' THEN T$(I)=SPACE(40-LENGTH(T$(I))/2)+T$(I)
1340     NEXT I
1350 K=6+3*(10-NS), N$=STR(NS), N1$=N$+'(BB %%.%)/'
1360 PRINT
1370 PRINT;;'0=MIN & MAX, 1=LIMITS - WHICH':
1380 INPUT FLAG
1390 IF FLAG THEN PRINT;;'SIGNIFICANCE LEVEL':
      ELSE 1430
1400 INPUT SIG
1410 A$="B' UL '"+N1$, B$="B' LL '"+N1$
1420 GOTO 1440
1430 A$="B'MAX '"+N1$, B$="B'MIN '"+N1$
1440 HED$='///'+STR(K)+"B'HOUR    '"+N$+"(6%)/"+STR(K+6)+
      'B '+N$+'(6%)/'+STR(K+7)+'B '+N$+"('   H+'%%)/"
1450 FMT$=STR(K+4)+"B'N '"+N$+'(6%)/'+STR(K+3)+A$+STR(K+3)+
      "B'MED '"+N$+'(BB %%.%)/'+STR(K+3)+B$
1460 TIME=100*FIX(.01*TIME)
1470 MAT CV=(-1)
1480 GOSUB 1850
1490     FOR I3=1 TO NR
1500     IF TIME=0 OR TIME=2500 THEN TIM$='0000'
          ELSE TIM$=STR(TIME)
```

```
1510     IF LENGTH(TIM$)=4 THEN PRINT TAB(K):TIM$
          ELSE PRINT TAB(K):'0':TIM$
1520     TIME=TIME+100
1530     MAT Y=(0)
1540        FOR I2=1 TO NS
1550        N=0
1560           FOR I1=1 TO ND
1570           IF X(I1,I3,I2)>=0 THEN N=N+1, SRT(N)=X(I1,I3,I2)
1580           NEXT I1
1590        Y(1,I2)=N, FAC1=8191
1591        IF N THEN 1600 ELSE Y(2,I2),Y(3,I2),Y(4,I2)=1E6
1592        GOTO 1740
1600           FOR J1=1 TO 12
1610           FAC1=(FAC1-1)/2
1620              FOR J2=FAC1+1 TO N
1630              TEMP=SRT(J2), FAC2=J2-FAC1
1640                 FOR J3=1 TO FAC2 STEP FAC1
1650                 FAC3=FAC2-J3+1
1660                 IF TEMP>=SRT(FAC3) THEN 1700                              EL
SE SRT(FAC3+FAC1)=SRT(FAC3)
1670                 NEXT J3
1680              SRT(FAC3)=TEMP
1690              GOTO 1710
1700              SRT(FAC3+FAC1)=TEMP
1710           NEXT J2,J1
1720       ( IF (N MOD 2) THEN Y(3,I2)=SRT((N+1)/2)
            ELSE Y(3,I2)=(SRT(N/2)+SRT(N/2+1))/2
1730        IF FLAG THEN GOSUB 1910            ELSE Y(2,I2)=SRT(N), Y(4,
I2)=SRT(1)
1740        NEXT I2
1750     MAT PRINT IN FORM FMT$:Y
1760     IF NOT (I3 MOD 8) THEN GOSUB 1830
1770     NEXT I3
1780 DO 1500: 1510
1790 PRINT CHAR(12)
1800 END
1810 PRINT;;'NOT ENOUGH DATA IN FILE ':F$
1820 END
1830 DO 1500: 1510
1840 T$(1)=SUBSTR(T$(1),6)+' (CONTINUED)'
1850 PRINT IN FORM'%//':CHAR(12)
1860     FOR I=1 TO 6
1870     IF T$(I)='0' THEN PRINT ELSE PRINT T$(I)
1880     NEXT I
1890 MAT PRINT IN FORM HED$:S$,H$
1900 RETURN
1910 IF CV(N)>=0 THEN 1940       ELSE NN=SIG/(.5^(N-1)), A,A1=1, I=0
1920 I=I+1, A1=A1*(N+1-I)/I, A=A+A1
1930 IF A<NN THEN 1920 ELSE CV(N)=I-1
1940 Y(2,I2)=SRT(N-CV(N)), Y(4,I2)=SRT(CV(N)+1)
1950 RETURN
1960 PRINT,'***** PROGRAM 5;LSTAT;SUMMARY *****'
1970 PRINT
1980 PRINT;;;'THIS PROGRAM IS INTENDED PRIMARILY FOR SUMMARIZING'
1990 PRINT;;"ENVIRONMENTAL DATA FROM SEVERAL APCD'S AND/OR SITES IN"
2000 PRINT;;'A PARTICULAR GENERAL LOCATION.  READINGS ARE ASSUMED TO BE
'
 2010 PRINT;;'OBTAINED HOURLY AT EACH SITE - PROVISION IS MADE FOR MISSI
NG'
2020 PRINT;;'READINGS.'
```

```
2030 PRINT
2040 PRINT;;;'BEFORE RUNNING THIS PROGRAM, PREPARE A SEQUENTIAL TEXT DA
TA'
2050 PRINT;;'FILE IN THE FOLLOWING FORMAT:'
2060 PRINT
2070 PRINT,'X(1,1,1),X(1,1,2),....,X(1,1,S)'
2080 PRINT,'X(1,2,1),X(1,2,2),....,X(1,2,S)'
2090 PRINT,'.............................'
2100 PRINT,'X(1,R,1),X(1,R,2),....,X(1,R,S)'
2110 PRINT,'X(2,1,1),X(2,1,2),....,X(2,1,S)'
2120 PRINT,'.............................'
2130 PRINT,'X(D,R,1),X(D,R,2),....,X(D,R,S)'
2140 PRINT
2150 PRINT;;;'WHERE X(I,J,K) = DATA AS FOLLOWS:'
2160 PRINT,'I = DAY (I = 1 TO NUMBER OF DAYS, D)'
2165 PRINT,'J = READING TIME (J = 1 TO NUMBER OF READINGS/DAY, R)'
2170 PRINT,"K = SITE (K = 1 TO NUMBER OF SITES+APCD'S, S)"
2190 PRINT
2200 PRINT;;;'*** NOTE *** IF NO READING WAS OBTAINED AT A GIVEN'
2210 PRINT;;"TIME, ENTER '-1' IN THE FILE MATRIX TO INDICATE NO READING
."
2220 PRINT
2230 PRINT;;;'UPON LINKING THIS PROGRAM, THE USER WILL BE REQUESTED'
2240 PRINT;;"TO INPUT THE DATA FILENAME, THE NUMBER OF APCD'S, THE NUMB
ER"
2250 PRINT;;'OF SITES, THE NUMBER OF DAYS READINGS WERE TAKEN, AND THE'
2260 PRINT;;'AND THE NUMBER OF READINGS THAT WERE TAKEN EACH DAY.'
2270 PRINT
2280 PRINT;;;'THE USER IS THEN REQUESTED TO INPUT THE (MILITARY) TIME'
2290 PRINT;;'OF THE FIRST READING (I.E. 0700 - USE EARLIEST FULL HOUR)'
2300 PRINT;;'AND AN IDENTIFICATION SYMBOL AND NUMBER OF MINUTES AFTER'
2310 PRINT;;'THE HOUR READ FOR EACH APCD AND/OR SITE.'
2320 PRINT
2330 PRINT;;;'THE USER WILL THEN BE ASKED TO INPUT 6 LINES FOR A'
2340 PRINT;;'HEADING WHICH WILL BE CENTERED ON AN 8-1/2 INCH PAGE.'
2350 PRINT
2360 PRINT;;;'FINALLY, THE USER WILL BE ASKED IF HE WISHES THE MINIMUM'
2370 PRINT;;'AND MAXIMUM VALUES OR CONFIDENCE LIMITS TO BE PRINTED IN T
HE'
2380 PRINT;;'SUMMARY.  IF CONFIDENCE LIMITS ARE REQUESTED, THE USER WIL
L'
2390 PRINT;;'BE REQUESTED TO INPUT A SIGNIFICANCE LEVEL.'
2400 PRINT
2410 PRINT;;'OUTPUT CONSISTS OF A SUMMARY TABLE CONTAINING THE FOLLOWIN
G'
2420 PRINT;;'HOURLY ENTRIES FOR EACH APCD AND SITE AS IDENTIFIED BY INP
UT:'
2430 PRINT
2440 PRINT,'NUMBER OF OBSERVATIONS, N'
2450 PRINT,'EITHER THE UPPER CONFIDENCE LIMIT, UL OR MAXIMUM VALUE, MAX
'
 2460 PRINT,'THE MEDIAN VALUE, MED'
2470 PRINT,'EITHER THE LOWER CONFIDENCE LIMIT, LL OR MINIMUM VALUE, MIN
'
```

5;LAB;LARSEN

```
COPY %LARSEN TO TEL TEXT  2/04/75
10 PRINT CHAR(12)
20 PRINT
30 PRINT IN FORM "35B 10% / ":TDATE
40 PRINT
50 PRINT "          5,LAB,LARSEN--FOR DATA STATISTICS AND/OR LARSEN'S MOD
EL"
51 PRINT "          LATEST REVISION - 1/31/75"
60 PRINT
70 PRINT "          QUESTIONS? CALL PAUL ALLEN 8-432-4877"
80 PRINT
90 GOTO 110
100 LINK "5,LAB,LAREXP"
110 PRINT "          PROGRAM EXPLANATION (YES OR NO) ":
120 INPUT DOY$
130 IF DOY$ = "YES" THEN 100 ELSE 140
140 PRINT
150 PRINT
160 PRINT "          JOB TITLE = ":
170 INPUT DES$
180 PRINT
190 PRINT "          ARE YOU INPUTTING A DATA FILE (YES OR NO) ":
195 UN$ = 'N'
200 INPUT A$
210 IF A$ = "YES" THEN 230
220 IF A$ = "NO" THEN 2180 ELSE 190
230 PRINT
240 PRINT "          FILE NAME = ":
250 INPUT FIL$
260 PRINT
270 PRINT "          NUMBER OF ROWS IN YOUR FILE =":
280 INPUT II
290 PRINT
300 PRINT "          NUMBER OF COLUMNS IN YOUR FILE =":
310 INPUT T
320 PRINT
330 PRINT
340 PRINT "          HOW MANY COLUMNS TO BE INCLUDED IN THIS RUN ":
350 INPUT JJ
360 PRINT
370 OPEN FIL$,1,SYMBOLIC,INPUT,SEQUENTIAL,OLD
380 ON ENDFILE 1 GOTO 410
390 DIM N(II,T), M(JJ), TOT(2)
400 MAT INPUT FROM 1: N
410 CLOSE 1
420 IF JJ = T THEN 430 ELSE 470
430 FOR L = 1 TO T
440 M(L) = L
450 NEXT L
460 GOTO 510
470 PRINT "          COLUMN NUMBERS TO BE USED IN THIS RUN ARE?":
480 FOR L= 1 TO JJ
490 INPUT M(L)
500 NEXT L
510 PRINT
520 PRINT "          INTERVAL FOR DATA SEARCH (CELL INTERVAL) = ":
530 INPUT FGW
540 PRINT
550 SUM=0
560 NEG =0
```

```
570 FOR L= 1 TO JJ
580 A = M(L)
590 FOR K = 1 TO II
600 IF N(K,A) = -1 THEN 630
610 SUM = N(K,A) + SUM
620 GOTO 640
630 NEG = NEG + 1
640 NEXT K
650 NEXT L
660 MP = SUM/(II*JJ-NEG)
670 DIFF = 0
680 FOR L = 1 TO JJ
690 A = M(L)
700 FOR K = 1 TO II
710 IF N(K,A) = -1 THEN 730
720 DIFF = DIFF + (N(K,A) - MP)**2
730 NEXT K
740 NEXT L
750 SD = SQRT(DIFF/(II*JJ-NEG-1))
760 SUM = 0
770 D = MP - SD/2
780 B = MP + SD/2
790 FOR L = 1 TO JJ
800 A = M(L)
810 FOR K = 1 TO II
820 IF N(K,A) = -1 THEN 850
830 IF N(K,A)>B THEN B =N(K,A)
840 IF N(K,A)< D THEN D = N(K,A)
850 NEXT K
860 NEXT L
870 PRINT "         DO YOU WISH DATA LISTING (YES OR NO) ":
880 INPUT AN$
890 IF AN$ = "YES" THEN 910
900 IF AN$ = "NO" THEN 1200 ELSE 870
910 PRINT CHAR(12)
920 PRINT CHAR(10)
930 PRINT IN FORM "35B 10% /":TDATE
940 PRINT
950 PRINT SPACE(40-ROUND(.5*(LENGTH(DES$)))):DES$
960 PRINT CHAR (10)
970 PRINT "       DATA"
980 PRINT CHAR (10)
990 FOR C = 1 TO JJ
1000 PRINT TAB(6*C +9):M(C):
1010 NEXT C
1020 PRINT
1030 PDA = 5+6*JJ
1040 FOR$ = ''
1050 FOR  I = 1 TO PDA
1060 FOR$ = FOR$ + '*'
1070 NEXT I
1080 FOR$ = '        ' + FOR$
1090 FOR$ = "'" + FOR$ + "'"
1100 PRINT IN FORM FOR$:
1110 PRINT CHAR (10)
1120 FOR K = 1 TO II
1130 PRINT IN FORM "4B 4% '--' BB ":K
1140 FOR L = 1 TO JJ
1150 A = M(L)
1160 PRINT IN FORM "%%.%% B":N(K,A)
```

```
1170 NEXT L
1180 PRINT
1190 NEXT K
1200 PRINT CHAR(12)
1210 PRINT CHAR(10)
1220 PRINT CHAR(10)
1230 PRINT IN FORM "35B 10% /":TDATE
1240 PRINT
1250 PRINT SPACE(40-ROUND(.5*(LENGTH(DES$)))):DES$
1260 PRINT CHAR (10)
1270 PRINT IN FORM "6B 'ARITHMETIC MEAN = ' %%%.%%% /":MP
1280 PRINT IN FORM "6B 'STANDARD ARITHMETIC DEVIATION = ' %%%.%%% /":SD
1290 PRINT
1300 PRINT IN FORM "6B 8% BB 4%.%% BB %% 6%.%% /":"RANGE IS",D,"TO",B
1310 IF YQ=2 THEN PRINT IN FORM "6B 'MEDIAN = ' 4%.2% ":TEST
1315 IF YQ = 2 THEN 1460
1320 TEST = D
1330 MM = 0
1340 FOR L = 1 TO JJ
1350 A = M(L)
1360 FOR K = 1 TO II
1370 IF N(K,A) = -1 THEN 1400
1380 IF N(K,A) < TEST THEN  MM = MM+ 1
1390 IF N(K,A) = TEST THEN MM = MM + 1
1400 NEXT K
1410 NEXT L
1420 IF MM>(II*JJ-NEG)/2 THEN PRINT IN FORM "6B 'MEDIAN = ' 4%.2% ":TES
T
1430 IF MM > (II*JJ-NEG)/2 THEN 1460
1440 TEST = TEST + FGW
1450 GOTO 1330
1460 PRINT CHAR(10)
1470 IF YQ = 2 THEN PRINT "      SAMPLE SIZE = ":SAMP
1480 IF YQ = 2 THEN GOTO 1500
1490 PRINT "      SAMPLE SIZE = ":II*JJ-NEG
1500 PRINT CHAR(10)
1510 PRINT IN FORM "31B 22%":"FREQUENCY DISTRIBUTION"
1520 PRINT CHAR (10)
1530 ROC$ = "16B 12% 5B 10% 5B 11% 5B 9%"
1540 PRINT IN FORM ROC$:"FREQUENCY OF","CUMULATIVE","  LARSEN'S ","NUMB
ER OF"
1550 PRINT
1560 OCC$ = "17B 10% 6B 9% 6B 11% 4B 11% "
1570 PRINT IN FORM OCC$:"OCCURRENCE","FREQUENCY"," FREQUENCY ","OCCURRE
NCES"
1571 PRINT IN FORM "20B '(%)' 12B '% =OR<' 10B '% =OR>' ":
1580 PRINT
1590 PRINT
1600 IF YQ = 2 THEN 2770
1610 MAT N = (1/FGW) * N
1620 REM    MAT N IS NOW AN ARRAY WITH A SPACING OF 1 BETWEEN SEARCHES.
               THIS ALLOWS USER TO COMPUTE STATISTICS FOR VARIOUS
1630 REM        INTERVAL SIZES FOR DISPLAYING ANY DATA BY HISTOGRAMS OR F
REQUENCY               PLOTS
1640 SSSS = (II*JJ)-NEG
1650 FOR TEST = ROUND(D/FGW) TO ROUND(B/FGW) STEP 1
1660 PCUM,CRNK=0
1670 FOR L = 1 TO JJ
1680 A = M(L)
```

```
1690 FOR K = 1 TO II
1700 IF N(K,A) < 0 THEN 1720
1710 IF N(K,A) < TEST + FGW/10 THEN PCUM = PCUM + 1
1720 NEXT K
1730 NEXT L
1740 PCUM = PCUM/((II*JJ)-NEG)
1750 IF PCI = PCUM THEN 1900 ELSE 1760
1760 PRINT IN FORM "4B 4%.2% '--' ":(TEST*FGW)
1770 T = PCUM-PCI
1780 BILL = ROUND(T*SSSS)
1790 TOM = TOM + BILL
1800 FOR I = (SSSS-TOM+1) TO (SSSS-TOM+BILL) STEP 1
1810 CRNK = CRNK + I
1820 NEXT I
1830 MRNK = CRNK/BILL
1840 CUMP = (MRNK-.4)/SSSS
1850 REM BILL = NUMBER OF OBSERVATIONS FOR VALUE TEST IN DATAFILE
             TOM = CUMULATIVE NUMBER OF OBSERVATIONS
1860 COR$ = "P18 %%.%%% P34 %%%.%%% P51 %%.%%% P64 %%%%%.%"
1867 CUMP = CUMP * 100.
1868 PPUM = PCUM * 100.
1869 T = T * 100.
1870 PRINT IN FORM COR$:T,PPUM,CUMP,BILL
1880 PRINT
1890 PCI = PCUM
1900 NEXT TEST
1910 DO 1500, 1500
1920 GOSUB 4980
1930 SGB = EXP(I-SQRT(I**2-2*LOG(B/MP)))
1940 MGB = B/(SGB^I)
1950 PRINT "        FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEAN: (*)"

1960 PRINT
1970 PRINT IN FORM "10B 'STANDARD GEOMETRIC DEVIATION = ' %%.%%% / ":SG
B
1980 PRINT IN FORM "10B 'GEOMETRIC MEAN = ' %%.%%% /":MGB
1990 PRINT CHAR(10)
2000 SGBC = EXP(SQRT(LOG(SD**2/MP**2+1)))
2010 MGBC = MP/EXP(0.5*LOG(SGBC)**2)
2020 PRINT "        FROM ALL OBSERVED DATA:"
2030 PRINT
2040 PRINT IN FORM "10B 'STANDARD GEOMETRIC DEVIATION = ' %%.%%% /":SGB
C
2050 PRINT IN FORM "10B 'GEOMETRIC MEAN = ' %%.%%  /":MGBC
2060 PRINT
2070 PRINT "        (*) REFERENCE AP-89"
2080 PRINT CHAR(12)
2090 PRINT
2100 PRINT "           USE LARSEN'S FREQUENCY FOR PLOTTING ON LOG-PROBABIL
ITY PAPER."
2110 PRINT CHAR(10)
2120 PRINT "           DO YOU WANT A LARSEN'S MODEL ANALYSIS"
2130 PRINT "           ON YOUR DATAFILE AT THIS TIME(YES OR NO)":
2140 INPUT AA$
2150 PRINT CHAR(10)
2160 IF AA$ = "YES" THEN GOTO 2790
2170 IF AA$ = "NO" THEN GOTO 5190 ELSE 2120
2180 PRINT
2190 PRINT "           DO YOU HAVE DISTRIBUTION STATS(1) OR A FREQ DIST.(2
)":
```

```
2200 INPUT YQ
2210 PRINT
2220 IF YQ = 2 THEN 2380
2230 IF YQ = 1 THEN 2240 ELSE 2190
2240 DIM FR(1)
2250 L99 = 1
2260 PRINT "          INPUT MAXIMUM OBSERVED CONCENTRATION":
2270 INPUT FR(L99)
2271 B = FR(L99)
2280 PRINT
2290 PRINT "          INPUT NUMBER OF OCCURRENCES OF THIS MAX.":
2300 INPUT NCM
2310 PRINT
2320 PRINT "          INPUT THE SAMPLE SIZE":
2330 INPUT SAMP
2340 PRINT
2350 PRINT "          INPUT THE ARITHMETIC MEAN":
2360 INPUT MP
2361 PRINT
2362 DO 520:540
2370 GOTO 3100
2380 PRINT CHAR(10)
2390 PRINT "          INPUT THE NUMBER OF INTERVALS IN THE DISTRIBUTION":

2400 INPUT L99
2410 PRINT
2420 DIM FR(L99), PFR(L99), CYMP(L99)
2430 PRINT "          ENTER THE CONCENTRATION AND ITS FREQUENCY (LO TO HI
) "
2440 PRINT "          (ENTER FREQUENCIES IN PERCENTAGES; I.E. TYPE 35 FOR
"
2450 PRINT "          35% - SUM OF ALL FREQS SHOULD APPROXIMATE 100.)"
2460 PRINT
2470 QDN$ = "FOR INTERVAL NO. "
2480 PND$ = "8B 17% B 3% "
2490 FOR J = 1 TO L99
2500 PRINT IN FORM PND$:QDN$,J
2510 INPUT FR(J),PFR(J)
2520 PRINT
2530 IF J = 1 THEN GOTO 2550
2540 IF FR(J) < FR(J-1) THEN 2430
2550 NEXT J
2551 SIUM = 0
2552 FOR I = 1 TO L99
2553 SIUM = SIUM + PFR(I)
2554 NEXT I
2555 IF SIUM>102. THEN 2557
2556 IF SIUM<98. THEN 2557 ELSE 2560
2557 PRINT IN FORM "8B 'THE SUM OF YOUR FREQUENCIES = ' %%%%.%% '%' /":
SIUM
2558 PRINT "          THIS SUM MUST BE WITHIN 2% , TRY AGAIN."
2559 GOTO 2430
2560 PRINT "          INPUT THE SAMPLE SIZE FOR THE DISTRIBUTION":
2570 INPUT SAMP
2571 DO 520:540
2580 PRINT
2590 MP = 0
2600 FOR J = 1 TO L99
2610 MP = MP + FR(J) * PFR(J)/100
2620 NEXT J
```

```
2630 SUM = 0
2640 FOR J = 1 TO L99
2650 SUM = SUM + (FR(J)**2) * PFR(J)/100
2660 NEXT J
2670 D = FR(1)
2680 B = FR(L99),CYMP(1),CYMP(2) = 0
2690 FOR J = 2 TO L99+1
2700 CYMP(J) = CYMP(J-1) + PFR(J-1)
2710 IF CYMP(J) < 50. THEN 2740
2720 IF CYMP(J)-50. < 50.-CYMP(J-1) THEN TEST = FR(J-1) ELSE TEST=FR(J-
2)
2730 GOTO 2750
2740 NEXT J
2750 SD = SQRT(SAMP/(SAMP-1) * (SUM -MP**2))
2760 GOTO 1200
2770 GOSUB 5050
2780 GOTO 3090
2790 SAMP = II*JJ-NEG
2791 PRINT
2800 PRINT "         DO YOU WANT A FREQUENCY LISTED FOR EACH VALUE IN TH
E"
2810 PRINT "         LARSEN DISTRIBUTION (YES OR NO) ":
2820 INPUT FLL$
2821 PRINT
2830 IF FLL$ = 'YES'THEN 2850
2840 IF FLL$ # 'NO' THEN 2800
2841 PRINT
2850 PRINT "         DO YOU WANT AN ANALYSIS FOR 1-HR AVE. TIME ":
2860 INPUT TT$
2861 PRINT
2870 IF TT$ = "YES" THEN 2880 ELSE 2900
2880 PRINT "         INPUT STANDARD FOR 1-HR AVE. TIME (SAME UNITS AS DA
TA)":
2890 INPUT OHST
2900 PRINT
2910 DO 3110, 3120
2920 PRINT
2930 PRINT "         HOW MANY OTHER AVERAGING TIMES DO YOU WANT ANALYZED
 ":
2940 INPUT N98
2950 IF N98 = 0 THEN 3035
2960 PRINT
2970 DIM ATS(N98),C99(N98)
2980 PRINT "         INPUT THE AVERAGING TIMES (HOURS) ":
2990 MAT INPUT ATS
3000 PRINT
3010 PRINT "         INPUT THE RESPECTIVE STANDARDS(SAME UNITS AS DATA)"
:
3020 MAT INPUT C99
3030 PRINT
3035 PRINT
3040 PRINT "         WHAT SAMPLE SIZE FOR 1-HR AVE. TIME DO YOU WANT"
3050 PRINT "         YOUR SAMPLE EXPANDED TO (HOURS) ":
3060 INPUT OHSS
3065 IF N98 = 0 THEN 3108
3070 IF TT$ = "NO" THEN 3230
3080 GOTO 3130
3090 PRINT
3095 DO 1910:2150
3096 IF AA$ = 'NO' THEN 5190
```

```
3100 DO 2800: 2900, 2920: 3070
3108 IF UN$ = 'PPM' THEN 3130
3109 IF UN$ = 'PPHM' THEN 3130
3110 PRINT "          WHAT UNITS ARE YOUR DATA IN(PPM OR PPHM)":
3120 INPUT UN$
3130 PRINT CHAR(12)
3140 PRINT CHAR(10)
3150 PRINT CHAR(10)
3160 PRINT IN FORM "35B 10% /":TDATE
3170 PRINT
3180 PRINT SPACE(40-ROUND(.5*(LENGTH(DES$)))):DES$
3190 PRINT CHAR(10)
3200 MUT$ = 'STANDARD GEOMETRIC DEVIATION = '
3210 PIR = 1/SQRT(2*PI)
3220 EAT$ = 'FROM SAMPLING PARAMETERS = '
3230 IF A$ = "YES" THEN 3290
3240 GOSUB 4770
3250 SGB = EXP(I - SQRT(I**2 - 2*LOG(FR(L99)/MP)))
3260 IF YQ # 1 THEN 3280
3270 SD = MP*(SQRT(EXP((LOG(SGB))^2)-1))
3280 MGB = B/(SGB^I)
3290 PRINT
3300 FREQ = .6/OHSS
3310 GOSUB 4570
3320 C = MGB*SGB**Z
3330 IF TT$ = "NO" THEN 3770
3340 PRINT "                   STATISTICAL PARAMETERS FROM 1-HOUR AVERAGING
TIME DATA"
3350 PRINT
3360 PRINT IN FORM "23B 'ARITHMETIC MEAN = ' %%.%%% /":MP
3370 PRINT IN FORM "23B 'STANDARD DEVIATION = ' %%.%%% /":SD
3380 PRINT
3390 PRINT IN FORM "19B 'FROM MAXIMUM OBSERVED VALUE AND ARITHMETIC MEA
N:' ":
3400 PRINT CHAR(10)
3410 PRINT IN FORM "23B 'STANDARD GEOMETRIC DEVIATION = ' %%.%%% /":SGB

3420 PRINT IN FORM "23B 'GEOMETRIC MEAN = ' %%.%%% /":MGB
3430 PRINT CHAR (10)
3440 PRINT "                         EXPECTED MAXIMUM 1-HOUR CONCENTRATION"
3450 PRINT IN FORM "19B 'FROM SAMPLING PARAMETERS = ' %%.%% B 4% /":C,U
N$
3460 PRINT
3470 PRINT IN FORM "19B 'NUMBER OF TIMES AMBIENT AIR QUALITY'/":
3480 I1 = (LOG(OHST) - LOG(MGB))/LOG(SGB)
3490 GOSUB 4400
3500 N35X = ROUND(OHSS*FREQ)
3510 PRINT IN FORM "19B 'STANDARD OF' B %%%.%% B 4% B 'TO BE EXCEEDED =
 ' 4%       B 'PER YEAR'/":OHST,UN$,N35X
3520 PRINT CHAR(10)
3530 IF FLL$ = 'NO' THEN 3740
3540 PRINT IN FORM "14B 'PREDICTED LARSENS FREQUENCIES FOR 1-HOUR CONCE
NTRATIONS' //":
3550 PRINT IN FORM "20B 'CONCENTRATION' 14B 'LARSENS FREQUENCY' /":
3560 PRINT IN FORM "24B %%%% 24B '% =OR>' /":UN$
3570 PRINT
3575 LOOP = ROUND(C/FGW)*FGW
3580 FOR J = FGW TO LOOP STEP FGW
3590 I1 = (LOG(J) - LOG(MGB))/LOG(SGB)
3610 GOSUB 4400
```

```
3619 F = F*100.
3620 PRINT IN FORM "22B %%.%% 26B %%.%%%% /":J,F
3720 NEXT J
3730 PRINT CHAR(10)
3740 PRINT IN FORM "13B 'SAMPLE SIZE =' B 4%//":OHSS
3750 PRINT "                ALL CALCULATIONS REFERENCE AP-89"
3760 PRINT
3770 PRINT CHAR(12)
3780 PRINT
3790 IF N98 = 0 THEN 5190
3800 FOR K = 1 TO N98
3810 GOSUB 4860
3820 CX = (C*(ATS(K)**Q))
3830 PRINT IN FORM "35B 10% /":TDATE
3840 PRINT
3850 DO 3180
3860 V = LOG(OHSS/ATS(K))/LOG(OHSS)
3870 BOM$ = "STATISTICAL PARAMETERS FROM"
3880 COM$ = "-HOUR AVERAGING TIME DATA"
3890 FOM$ = "12B 27% B %% 25% /"
3900 SGB1 = SGB ** SQRT(V)
3910 PRINT CHAR(10)
3920 PRINT IN FORM FOM$:BOM$,ATS(K),COM$
3930 PRINT
3940 MGB1 = MGB * SGB ** (0.5 * (1-V) * LOG(SGB)**2)
3950 PRINT IN FORM "23B 'GEOMETRIC MEAN = ' %%.%%% /":MGB1
3960 PRINT
3970 PRINT IN FORM "23B 'STANDARD GEOMETRIC DEVIATION = ' %%.%%% /":SGB
1
3980 PRINT CHAR(10)
3990 I1 = (LOG(C99(K)) - LOG(MGB1))/LOG(SGB1)
4000 GOG$ = "EXPECTED MAXIMUM"
4010 GOT$ = "-HOUR CONCENTRATION"
4020 HGO$ = "19B 16% B 4% 19% /"
4030 PRINT IN FORM HGO$:GOG$,ATS(K),GOT$
4040 PRINT IN FORM "19B 'FROM SAMPLING PARAMETERS = ' %%.%% B 4% /":CX,
UN$
4050 PRINT CHAR(10)
4060 GOSUB 4400
4070 DO 3470
4080 NATX = ROUND((OHSS/(ATS(K)))*FREQ)
4090 PRINT IN FORM "19B 'STANDARD OF ' 2%.2% B 4% ' TO BE EXCEEDED ='
     B 4% B 'PER YEAR'/":C99(K),UN$,NATX
4100 PRINT CHAR (10)
4110 IF FLL$ = 'NO' THEN 4310
4120 PRINT IN FORM "14B 'PREDICTED LARSENS FREQUENCIES FOR ' 4% '-HOUR
CONCENTRATIONS' //":ATS(K)
4125 DO 3550
4130 DO 3560
4140 PRINT
4145 LOOP = ROUND(CX/FGW)*FGW
4150 FOR J = FGW TO LOOP STEP FGW
4160 I1 = (LOG(J) - LOG(MGB1))/LOG(SGB1)
4180 GOSUB 4400
4189 F = F*100.
4190 PRINT IN FORM "22B %%.%% 26B %%.%%%% /":J,F
4290 NEXT J
4300 PRINT CHAR(10)
4310 PRINT IN FORM "12B 'SAMPLE SIZE = ' 4% /":ROUND(OHSS/(ATS(K)))
4320 PRINT CHAR(10)
```

```
4330 PRINT "                ALL CALCULATIONS REFERENCE AP-89"
4340 PRINT CHAR(12)
4350 NEXT K
4360 PRINT
4370 PRINT CHAR(10)
4380 GOTO 5190
4390 REM THIS IS BEGINNING OF SUBROUTINES
4399 REM  SUBROUTINE FROM LARSEN TO CALCULATE AN F FROM A Z
4400 Z = I1
4401 DEF FNZ2FR(F)
4402 DOUBLE F,Z
4410 TT= 1/(1+0.2316419*F)
4420 WW = 0.3989422804*(EXP(-0.5*F*F))
4430 Q = 1 - WW*(0.31938153*TT-0.356563782*TT*TT+1.781477937*TT^3
        -1.821255978*TT^4+1.330274429*TT^5)
4450 RETURN Q
4455 END
4460 SNZ = SGN(Z)+2
4470 ON SNZ GOTO 4500,4480,4520
4480 F = 0.5
4490 GOTO 4530
4500 F = FNZ2FR(ABS(Z))
4510 GOTO 4530
4520 F = 1 - FNZ2FR(Z)
4530 FREQ = F
4531 RETURN
4570 REM  SUB   FREQ TO GET Z
4580 F = FREQ
4590 SNF = SGN(F-.5)+2
4600 DEF FNZ(F)
4610 DOUBLE F
4620 DOUBLE Z
4630 WF = LOG(1/(F^2))^.5
4640 Z = WF-(2.515517+0.802853*WF+0.010328*WF^2)/
       (1+1.432788*WF+0.189269*WF^2+0.001308*WF^3)
4650 RETURN Z
4660 END
4670 ON SNF GOTO 4680, 4730, 4750
4680 IF F=0 THEN 4710
4690 Z = FNZ(F)
4700 GOTO 4760
4710 PRINT "ANS DEVIA = +/- INFINITY , PROB = ":F
4720 GOTO 4760
4730 Z = 0
4740 GOTO 4760
4750 Z = -(FNZ(1-F))
4760 RETURN
4770 REM   CALCULATION OF Z FOR MAX ABSERVED VALUE
4780 AREA = 0.5
4790 IF YQ = 1 THEN 4810
4800 NCM = ROUND(PFR(L99) *SAMP/100)
4810 RK = NCM * 0.5 + .5
4820 F = (RK -.4)/SAMP
4830 GOSUB 4590
4840 I = Z
4850 RETURN
4860 REM   CALCULATION OF Q   SLOPE OF MAXIMUM CONCENTRATION LINE
                    FOR 1-HOUR AVERAGING TIME STANDARD
4870 IF SGB>4.99 THEN Q=-.59
4880 IF SGB> 4.99 THEN RETURN
```

```
4890 DIM QQ(400)
4900 SGGB = 100 * SGB
4910 SSGB = ROUND(SGGB) + 1
4920 E = SSGB-100
4930 OPEN "5;LAB;QFIL",1,SYMBOLIC,INPUT,SEQUENTIAL,OLD
4940 MAT INPUT FROM 1:QQ
4950 CLOSE 1
4960 Q = QQ(E)
4970 RETURN
4980 REM  SUBROUTINE TO CALCULATE Z FOR MAXIMUM OBSERVED VALUE
               IN DATA FILE ANALYSIS
4990 NOM = BILL
5000 RNK = BILL*.5 + .5
5001 IF YQ = 2 THEN F = (RNK - .4)/SAMP
5002 IF YQ = 2 THEN 5020
5010 F = (RNK - .4)/(II*JJ - NEG)
5020 GOSUB 4590
5030 I = Z
5040 RETURN
5050 REM  SUBROUTINE TO CALCULATE FREQS, CUM FREQS, LARSEN' FREQS,
               AND NUMBER OF OCCURRENCES FOR STATISTIC PAGE OF FREQ DIST
INPUT PLOT
5060 DO 1860
5070 PCUM = 0,CRNK=0
5080 FOR J = 1 TO L99
5085 CRNK = 0 ←—
5090 PRINT IN FORM "4B 4%.2% '--' ":FR(J)
5100 T = PFR(J)/100
5110 PCUM = PCUM + T
5120 BILL = ROUND(T*SAMP)
5130 TOM = TOM+BILL
5131 FOR I = (SAMP-TOM+1) TO (SAMP-TOM+BILL) STEP 1
5132 CRNK = CRNK + I
5133 NEXT I
5135 MRNK = CRNK/BILL
5140 CUMP =(MRNK - .4)/SAMP
5157 CUMP = CUMP * 100.
5158 PCUM = PCUM * 100.
5159 T = T * 100.
5160 PRINT IN FORM COR$:T,PCUM,CUMP,BILL
5161 PCUM = PCUM/100.
5162 T = T/100.
5165 PRINT
5170 NEXT J
5180 RETURN
5190 END
-
```